aerospace

MDPI

Article

Empirically Calibrated Multi-Fidelity Fusion with Conformal Prediction Intervals for Reliability Assessment of Aerospace Dormant Components

Shengpeng Zhang [1,2], Shuanglong Rong [2], Hao Li [2], Shuo Huang [2], Cheng-Wei Fei [3] and Baiyang Zheng [3,*]

[1] School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
[2] Aerospace Science and Industry Defense Technology Research Testing Center, Beijing 100854, China
[3] College of Intelligent Robotics and Advanced Manufacturing, Fudan University, Shanghai 200433, China; cwfei@fudan.edu.cn
* Correspondence: zhengby25@m.fudan.edu.cn

Abstract

Reliability prediction of aerospace dormant components requires fusing natural-storage observations at the operating temperature with accelerated-storage testing data at elevated temperatures. Existing scalar-weight fusion methods apply a global weight that cannot reflect the time-varying trustworthiness of the accelerated branch as Arrhenius extrapolation distance grows. Physics-based fusion propagates accelerated-test scatter through least squares but leaves the dominant error source—the degradation-model form itself—unaccounted for, and no method in either class verifies the coverage of its intervals. This paper proposes an empirically calibrated multi-fidelity fusion that selects a mechanism-specific natural-branch degradation model by the corrected Akaike information criterion and augments the accelerated-branch variance with an additive model-form term fitted from natural-storage residuals. This term turns the fusion weight into a continuous, time-varying diagnostic that detects Arrhenius misspecification from training data alone and falls back safely to the natural-only estimate. Prediction intervals are calibrated by split-conformal prediction on a disjoint simulated population, giving finite-sample, distribution-free coverage, and the remaining-storage-life interval follows from the band's first-passage time. On a 1000-run varying-truth simulation, the calibrated band attains 95.5% trajectory coverage at the narrowest band width among six methods; on the torsion-bar case, the fusion reaches a held-out RMSE of 0.045 N·m and a remaining-life interval of 10.4–12.6 years. The model-form variance ratio provides a single-number regime diagnostic across all cases.

Keywords: accelerated storage testing; natural-storage data fusion; Arrhenius extrapolation uncertainty; conformal prediction; first-passage-time reliability; remaining storage life

Academic Editor: Konstantinos Tserpes

Received: 16 May 2026
Revised: 14 June 2026
Accepted: 25 June 2026
Published: 30 June 2026

1. Introduction

Aerospace dormant components, such as missile sealed amplifiers, torsion springs in launch mechanisms, and battery cells in standby avionics, spend most of their service life unused but must perform on first activation after years of storage. A single misjudged component can ground a system or cause mission failure, and the cost of over-conservative replacement schedules across a fleet is large. Prognostics and health-management practice for aerospace structures and propulsion systems is now a mature engineering discipline with broad theory and practice [1,2], and the reliability evaluation of structurally and

electronically heterogeneous aerospace components has driven a sustained body of work in this field. Examples include landing-gear fatigue-life prediction in both physics-based [3,4] and machine-learning formulations [5], compressor-blade radial-deformation reliability under crack defects [6], integrated long-term-deterioration monitoring of aero-engines [7,8], high-cycle fatigue assessment of helicopter flange structures [9], thermoacoustic fatigue analysis of thin-walled connection structures [10], and belief-reliability formulations developed for aerospace electromagnetic relays in storage [11]. Reliability assessment for dormant components in particular combines two heterogeneous degradation datasets: The first is natural-storage observations recorded over the deployed life at the operating temperature, which are sparse and slow. The second is accelerated-storage testing data recorded at elevated temperatures over short test windows, which are dense and fast but require extrapolation through an Arrhenius model to the operating temperature. The general statistical framework for fitting and analyzing such accelerated-life data is well established [12]. The specific sub-problem addressed in this paper is the principled fusion of these two datasets into a remaining-storage-life prediction with a calibrated credible interval, suitable for risk-tiered maintenance decisions.

A first class of approaches addresses this fusion through a scalar weighting of the two data sources. Naive averaging assigns each source equal weight. The relative-entropy weighting of [13] computes the weight from the Kullback–Leibler divergence between source predictions on the training grid. The Bayesian calibration-factor framework of [14], recently adapted to aerospace test-data fusion in [15] for landing-gear retraction-angle reliability, fits a posterior on a scalar that multiplicatively calibrates the accelerated prediction against natural-storage data. Related Bayesian fusion formulations have combined accelerated degradation data with field condition-monitoring data [16], addressed measurement-error correction when two types of accelerated data are jointly available [17], and used Bayesian model averaging across competing degradation processes (inverse Gaussian, gamma) to hedge against model uncertainty [18]. These methods have become canonical in the Chinese aerospace-reliability literature because they extend naturally from single-source maximum-likelihood inference and admit closed-form posteriors. The structural limitation common to the class is that the fusion weight is a global scalar, fixed across the entire prediction horizon. The trustworthiness of the accelerated branch is not in fact constant. It depends on the Arrhenius extrapolation distance from the test temperatures to the operating temperature, and on whether the Arrhenius assumption itself holds at the operating point. A scalar weight cannot reflect this time-varying trustworthiness, and when the Arrhenius assumption is misspecified the scalar weight cannot detect the misspecification from the data, pulling the fused prediction toward a biased accelerated estimate.

A second class of approaches uses the degradation mechanism explicitly. Physics-based degradation models, in the Arrhenius, Eyring, or power-law families, predict the operating-temperature trajectory from accelerated-test coefficient estimates, with uncertainty propagated through ordinary least squares. Multi-fidelity Gaussian-process (MFGP) methods, which trace back to the auto-regressive fidelity coupling of [19] and are reviewed for aerospace applications in [20] on the Gaussian-process side of the broader Gaussian-process literature [21], combine a high-fidelity but expensive source with a low-fidelity but cheap source through a precision-weighted posterior. Recent aerospace examples include the surrogate-based aerodynamic optimization work of [22], the multi-fidelity Hamilton–Kriging model with adaptive infill sampling for experimental optimization of [23], the deep-reinforcement-learning-enhanced Kriging method for aero-engine structural reliability of [24], the multi-fidelity reliability-based design optimization framework with a local-update surrogate strategy of [25], adaptive reliability analysis using a collective learning strategy across fidelity levels [26], and a physics-regression-oriented multi-fidelity

formulation [27]. A complementary line of work models the degradation path itself as a stochastic process, such as a Wiener process [28–30]. The limitation specific to storage reliability is that the dominant source of model-form error in this setting is not within-source noise or coarse-mesh approximation but the extrapolation distance between accelerated-test conditions and the operating temperature. Existing physics-based and multi-fidelity-GP fusion absorbs this extrapolation error into a generic noise term or an OLS-propagated variance, neither of which reflects the actual at-operating-temperature accuracy. As a consequence, no method in this class produces a prediction interval on remaining storage life whose finite-sample validity has been established against out-of-sample truth.

The common shortcoming across the two classes is that fusion weight and predictive variance are determined from quantities that do not see the operating-temperature ground truth at all. This issue is part of a broader epistemic-uncertainty representation problem under sparse and structurally heterogeneous data [31], but it takes a specific form in the storage-fusion setting. The technical gap that follows is twofold: First, the accelerated-branch variance should be decomposed into a within-branch intrinsic-noise component and a separate extrapolation-error component. The extrapolation-error component should be calibrated empirically against natural-storage residuals at the operating temperature, not propagated from accelerated-test scatter alone. Second, the resulting time-varying fusion weight should determine how the credible interval widens at each time, so that the band's empirical coverage matches the nominal level across the prediction horizon.

The mathematical machinery to close this gap is available. Precision weighting provides a combination of two source predictions with explicit per-branch variance. The natural-storage residuals against the accelerated prediction provide a direct empirical reference for the at-operating-temperature error that no accelerated-only variance estimate can reach. Information-criterion selection within a mechanism-specific model family ties the natural-branch mean to the failure physics rather than to a single assumed form. Split-conformal prediction supplies a finite-sample, distribution-free validity layer that no fusion pipeline in this literature includes. What remains is to assemble these pieces into a single closed-form procedure, validate the resulting interval against out-of-sample truth, and characterize the regime in which the procedure improves on natural-only rather than merely matching it.

This paper proposes an empirically calibrated multi-fidelity fusion for aerospace dormant-component reliability assessment. The method addresses the gap identified above through three contributions: First, the natural-branch mean is selected by AICc within a mechanism-specific nested family covering the primary and secondary stages of the failure mechanism, and the accelerated-branch variance is decomposed into intrinsic, sampling, and model-form components, the last fitted additively from natural-storage residuals. This produces a continuous, time-varying fusion weight that detects Arrhenius misspecification from data with no thresholds anywhere in the pipeline, where scalar-weight baselines such as KL-entropy and Bayesian calibration factor cannot. Second, the fused band is calibrated by split-conformal prediction and propagated through a first-passage-time construction to a prediction interval on the remaining storage life. On a 1000-run varying-truth Monte Carlo population with disjoint calibration and test halves, the calibrated band attains 95.5% out-of-sample trajectory coverage at a mean width of 0.071 N·m, 97 to 196 times narrower than the other fusion strategies at the same guaranteed level. Third, the model-form variance ratio is established as a single-number diagnostic for the fusion regime. Values of approximately 11,600 on a torsion-bar case and 81 on a sealed-amplifier case identify severe Arrhenius misspecification in advance, allowing the engineer to predict whether fusion will deliver calibrated uncertainty on a fallback-to-natural-only estimate or precision-weighted point-estimate fusion. The remainder of this paper is organized as

follows: Section 2 formalizes the storage-fusion problem and derives the proposed method. Section 3 reports the experimental setup, baseline methods, and held-out validation on two real case studies and one synthetic Monte Carlo simulation. Section 4 interprets the findings and compares them with prior work. Section 5 concludes the paper.

2. Proposed Method

The proposed method addresses the gap identified in Section 1. Natural-storage and accelerated-storage data carry complementary information, but existing fusion methods conflate their uncertainty sources and produce prediction intervals whose coverage is never verified. Section 2.1 states the formal estimation problem and the degradation parameterization shared by both case studies. Section 2.2 develops the multi-fidelity fusion with decomposed accelerated-branch variance, a mechanism-specific natural-branch model selected by the corrected Akaike information criterion (AICc), and an additive empirical model-form variance fitted from natural-storage residuals. Section 2.3 constructs a split-conformal prediction interval on the remaining storage life from the fused trajectory's first-passage time. Section 2.4 consolidates the procedure into a single algorithm.

2.1. Problem Definition and Preliminaries

A monitored component is stored dormant at operating temperature T_{op} for a period of years. Two data streams are available. The natural-storage stream $\mathbf{Y}^{(n)}$ records the degradation parameter $P(t)$ at coarse intervals over the full storage horizon; one observation $Y_i^{(n)}$ is the mean across n_{spec} specimens at time t_i. The accelerated-storage stream $\mathbf{Y}^{(a,k)}$ records the same parameter at K elevated temperatures $T_k > T_{\mathrm{op}}$, at fine temporal resolution but over a short test window $[0, t_{\mathrm{acc}}]$, with n_{acc} specimens per temperature. The two streams therefore probe different regions of the (time, temperature) plane and have different sample and noise structures.

Both stress relaxation and electronic drift follow a degradation form

$$\frac{P_0 - P(t)}{P_0} = A(T) + B(T)\,\phi(t), \tag{1}$$

where P_0 is the initial value of the parameter, $A(T)$ is a temperature-dependent intercept, $B(T)$ is the rate coefficient, and $\phi(t)$ is the mechanism-specific primary time feature $\phi(t) = \ln t$ for stress relaxation, where logarithmic kinetics arise from the progressive exhaustion of dislocation glide [32], and $\phi(t) = t$ for electronic drift governed by steady charge-trapping kinetics [33]. Each mechanism also admits a secondary stage at long horizons: stress relaxation, a linear term associated with grain-boundary sliding and secondary creep [32]; and drift, a logarithmic early transient from fast trap filling [33]. Section 2.2 therefore equips the natural branch with the nested two-member family $\{\phi(t)\} \cup \{\phi(t), \psi(t)\}$, where ψ denotes the secondary complement, and selects between the two forms by AICc; the accelerated branch retains the primary term alone, because the secondary stage is not identifiable within the short accelerated window—a point Section 4 returns to. We adopt the Arrhenius assumption on $B(T)$,

$$\ln B(T) = \ln B_0 - \frac{E_a}{R\,T}, \tag{2}$$

where E_a is the apparent activation energy and R the gas constant. Failure is declared when $P(t)$ crosses a fixed threshold P_{fail}.

The estimation problem is to construct, from $\mathbf{Y}^{(n)}$ and $\{\mathbf{Y}^{(a,k)}\}_{k=1}^{K}$, a fused trajectory $\hat{P}(t)$ together with a predictive variance $\hat{\sigma}^2(t)$, both defined at the operating temperature T_{op}, such that two requirements hold: First, the point estimate $\hat{P}(t)$ is closer to the unknown

truth than any single-source estimator on a held-out window of the natural-storage data. Second, the $1-\alpha$ prediction interval on the remaining storage life implied by $\hat{P}(t)$ and $\hat{\sigma}^2(t)$ has out-of-sample empirical coverage at the nominal level when validated against disjoint profiles with known truth. Figure 1 previews the procedure that meets these two requirements; the three subsequent subsections develop its components in turn.

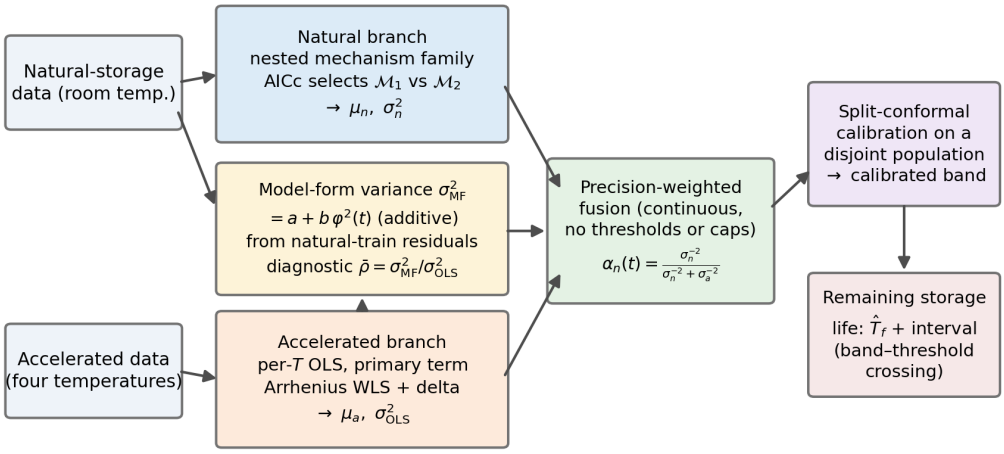


Figure 1. Overview of the proposed empirically calibrated multi-fidelity fusion with conformal prediction intervals. Two heterogeneous inputs feed two branches. The natural branch (top) fits the mechanism-specific nested family of Equation (9) and selects between the primary-only and primary-plus-secondary forms by AICc, returning μ_n and σ_n^2. The accelerated branch (bottom) runs per-temperature OLS on the primary term, Arrhenius WLS, and delta-method propagation, returning μ_a and the sampling variance $\sigma_{\mathrm{OLS}}^2(t)$. The additive model-form variance $\sigma_{\mathrm{MF}}^2(t) = a + b\,\varphi^2(t)$ (center), fitted from natural-training residuals, augments the accelerated variance and yields the regime diagnostic $\bar{\rho}$. Precision weighting fuses the branches through the continuous weight $\alpha_n(t)$ with no thresholds or caps. The fused band is calibrated by split-conformal prediction on a disjoint population, and the remaining-storage-life interval is read off the calibrated band's threshold crossings.

Naive fusion fails the second requirement for a specific structural reason. The accelerated branch reaches the natural-storage time horizon only through Arrhenius extrapolation across two orders of magnitude in time. The variance of this extrapolation grows with time through the $\phi(t)^2$ factor in the propagated regression covariance. Scalar fusion weights, of the kind used by KL-entropy and Bayesian calibration-factor methods, cannot reflect this time-varying trustworthiness. Pointwise weights without an honest treatment of model-form error cannot reflect that the Arrhenius model itself is misspecified at the extrapolation extreme. Section 2.2 addresses both issues.

2.2. Multi-Fidelity Fusion with Decomposed Accelerated-Branch Variance

The accelerated branch carries two distinct sources of uncertainty: specimen-to-specimen scatter at each test temperature produces an intrinsic noise floor; Arrhenius extrapolation from test temperatures to the operating temperature adds a separate, time-varying contribution. Existing fusion methods absorb both into a single residual variance and treat the accelerated branch as uniformly informative across the storage horizon. Two consequences follow: When Arrhenius extrapolation is reliable, the conflation under-uses the accelerated branch's actual time-varying precision. When Arrhenius extrapolation is misspecified at long distance, the conflation lets the accelerated branch silently bias the fused estimate. The first half of this subsection separates the two contributions and recovers a time-varying weight. The second half adds an additive empirical model-form variance, so that the time-varying weight also detects model-form misspecification and

shifts continuously to a safe natural-only fallback when the accelerated branch cannot be trusted.

The pipeline has three stages: First, at each accelerated temperature T_k, ordinary least squares regresses the normalized degradation $(P_0 - P)/P_0$ on the time feature $\phi(t)$, yielding coefficients $(\hat{A}_k, \hat{B}_k)$ together with standard errors and a per-temperature residual variance $\hat{\sigma}_k^2$. Second, weighted least squares on $\ln \hat{B}_k$ versus $1/T_k$ extracts the Arrhenius parameters and propagates the regression covariance to the operating temperature. Third, the accelerated prediction at any natural time t takes the form of Equation (1) with the room-temperature coefficients, and its variance decomposes into an intrinsic term and an extrapolation term.

At each accelerated temperature, the per-temperature regression is

$$\frac{P_0 - P}{P_0} = A_k + B_k \phi(t) + \varepsilon_k, \qquad \varepsilon_k \sim \mathcal{N}(0, \sigma_k^2). \tag{3}$$

The Arrhenius fit uses weighted least squares with weights derived from the standard errors $\mathrm{SE}(\ln \hat{B}_k)$:

$$\ln \hat{B}_{\mathrm{op}} = \gamma_0 + \gamma_1 / T_{\mathrm{op}}, \qquad \mathrm{Var}(\ln \hat{B}_{\mathrm{op}}) = \mathbf{x}_{\mathrm{op}}^{\top} \boldsymbol{\Sigma}_{\gamma} \mathbf{x}_{\mathrm{op}}, \tag{4}$$

where $\mathbf{x}_{\mathrm{op}} = (1, 1/T_{\mathrm{op}})^{\top}$ and $\boldsymbol{\Sigma}_{\gamma}$ is the WLS covariance of the slope-intercept estimate. An analogous WLS fit on $\ln \hat{A}_k$ yields $\hat{A}_{\mathrm{op}}$ and $\mathrm{Var}(\hat{A}_{\mathrm{op}})$. With $\hat{B}_{\mathrm{op}} = \exp(\ln \hat{B}_{\mathrm{op}})$ and the delta method providing $\mathrm{Var}(\hat{B}_{\mathrm{op}}) = \hat{B}_{\mathrm{op}}^2 \mathrm{Var}(\ln \hat{B}_{\mathrm{op}})$, the accelerated-branch variance at natural time t decomposes as

$$\sigma_a^2(t) = \sigma_{\mathrm{intrinsic}}^2 + \sigma_{\mathrm{extrap,OLS}}^2(t), \tag{5}$$

with $\sigma_{\mathrm{intrinsic}}^2 = \overline{\sigma_k^2}$ pooled over temperatures and $\sigma_{\mathrm{extrap,OLS}}^2(t) = \mathrm{Var}(\hat{A}_{\mathrm{op}}) + \mathrm{Var}(\hat{B}_{\mathrm{op}})\, \phi(t)^2$ derived from propagated regression covariance.

The decomposition makes the time dependence of the accelerated branch's trustworthiness visible. Intrinsic scatter is roughly constant in t; extrapolation variance grows like $\phi(t)^2$. Beyond the crossover time at which $\sigma_{\mathrm{extrap,OLS}}^2(t) = \sigma_{\mathrm{intrinsic}}^2$, Arrhenius extrapolation rather than measurement noise dominates the accelerated prediction's uncertainty. This diagnostic is invisible to scalar-weighting methods that collapse σ_a^2 into one number.

Propagated regression variance accounts only for sampling uncertainty in the Arrhenius fit, not for model-form error at the extrapolation extreme. For long-distance extrapolation, the Arrhenius assumption of Equation (2) and the family-specific time profile $\phi(t)$ together carry substantial model-form uncertainty that the OLS covariance misses. We therefore add a non-parametric model-form discrepancy variance, estimated from training-window residuals between the accelerated prediction and the natural-storage observation. Let $r(t_i) = Y_i^{(n)} - \hat{P}_a(t_i)$ denote these residuals on the training subset $\{t_i\}$ of natural-storage time points, where $\hat{P}_a(t)$ is the accelerated prediction at T_{op} from Equations (1) and (4). We fit

$$r^2(t) \approx a + b\, \varphi^2(t), \qquad a, b \geq 0, \tag{6}$$

where $\varphi(t)$ is the primary time feature in natural-branch units, and the coefficients are constrained non-negative. The φ^2 regressor is not an arbitrary choice; it mirrors the $\phi(t)^2$ growth of the propagated regression covariance in Equation (5), so that the empirical term and the OLS term share the same structural time dependence and differ only in scale. The model-form variance and the total accelerated-branch variance are then

$$\sigma_{\mathrm{MF}}^2(t) = \hat{a} + \hat{b}\, \varphi^2(t), \qquad \sigma_a^2(t) = \sigma_{\mathrm{intrinsic}}^2 + \sigma_{\mathrm{extrap,OLS}}^2(t) + \sigma_{\mathrm{MF}}^2(t). \tag{7}$$

The model-form term enters additively, as an independent variance source alongside sampling variance, so the total accelerated-branch variance is a smooth, continuous function of the data with no thresholds, caps, or switching rules anywhere in the pipeline. Where the Arrhenius assumption holds, the fitted $(\hat{a}, \hat{b})$ are small and the OLS term dominates; where it is misspecified, σ^2_{MF} grows and continuously down-weights the accelerated branch. The natural-storage training data plays a dual role here, both as a fusion input and as the residual reference for the model-form variance. This dual role is justified by the absence of any independent residual reference at the operating temperature. The implication is that the model-form variance is informative about which branch the fusion should trust, rather than an independent measure of accelerated-branch accuracy; for the same reason, interval calibration is never claimed from this in-sample construction but is instead established on fully disjoint, independently drawn profiles by the split-conformal procedure of Section 2.3. We refer to the time-averaged ratio

$$\bar{\rho} \;=\; \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{\sigma^2_{\mathrm{MF}}(t)}{\sigma^2_{\mathrm{extrap,OLS}}(t)} \tag{8}$$

over the prediction grid $\mathcal{T}$ as the model-form variance ratio; Section 4 uses $\bar{\rho}$ as a single-number diagnostic for the fusion regime.

The natural branch contributes a prediction $\hat{P}_n(t)$ and a variance $\hat{\sigma}^2_n(t)$. Both come from OLS regression on the natural-storage training data, with the mean model selected within the mechanism-specific nested family of Section 2.1,

$$\mathcal{M}_1 \colon \frac{P_0 - P(t)}{P_0} = \beta_0 + \beta_1 \varphi(t), \qquad \mathcal{M}_2 \colon \frac{P_0 - P(t)}{P_0} = \beta_0 + \beta_1 \varphi(t) + \beta_2 \psi(t), \tag{9}$$

by the small-sample corrected Akaike information criterion AICc [34,35]. The nested structure restricts selection to physically admissible forms for the failure mechanism at hand, rather than an open-ended basis search, and the selected design matrix supplies the standard prediction-variance inflation for extrapolation away from the training mean. The accelerated branch deliberately keeps the primary-only form $\mathcal{M}_1$: over a test window of days, the secondary term contributes at or below the accelerated noise floor and is not identifiable, which is precisely why the accelerated extrapolation can be biased at decade horizons and why σ^2_{MF} is needed. The two branches fuse by precision weighting,

$$\alpha_n(t) = \frac{1/\hat{\sigma}^2_n(t)}{1/\hat{\sigma}^2_n(t) + 1/\sigma^2_a(t)}, \tag{10}$$

$$\hat{P}(t) = \alpha_n(t)\, \hat{P}_n(t) + \big(1 - \alpha_n(t)\big)\, \hat{P}_a(t), \qquad \hat{\sigma}^2(t) = \frac{1}{1/\hat{\sigma}^2_n(t) + 1/\sigma^2_a(t)}. \tag{11}$$

Two limits clarify the design. When the model-form variance diagnoses severe Arrhenius misspecification, $\sigma^2_a(t)$ grows and $\alpha_n(t) \to 1$; the method degrades gracefully to a natural-only estimator. When natural-storage data is exhausted at long horizons, $\hat{\sigma}^2_n(t)$ grows under extrapolation and $\alpha_n(t)$ decreases, so the accelerated branch contributes more weight. Because σ^2_{MF} enters Equation (7) additively, $\alpha_n(t)$ is a continuous functional of the data throughout; the pipeline contains no hard thresholds, weight caps, or heuristic switching rules. The time-varying $\alpha_n(t)$ replaces the scalar weight that prior fusion methods use.

2.3. Conformal Prediction Interval on the Remaining Storage Life

A point estimate of the remaining storage life is of limited operational value without an honest interval. Decisions about when to inspect, replace, or retire a stored component

depend on the lower end of a prediction interval, not on the median prediction. Existing fusion methods either report point estimates or rely on intervals whose empirical coverage has not been validated. We calibrate the predictive band by split-conformal prediction [36–38], which carries a finite-sample, distribution-free coverage guarantee and requires no Gaussian assumption on the fused errors, and we read the remaining-storage-life interval off the calibrated band.

The fused trajectory $\hat{P}(t)$ with variance $\hat{\sigma}^2(t)$ defines a predictive band $\hat{P}(t) \pm \kappa\,\hat{\sigma}(t)$ whose half-width multiplier κ must be chosen. The Gaussian choice $\kappa = z_{\alpha/2}$ presumes normal fused errors; we instead calibrate κ by split-conformal prediction. Given a calibration set of profiles disjoint from any test profile, the standardized non-conformity scores are

$$s_j = \frac{\left|P_{\text{truth},j} - \hat{P}_j\right|}{\hat{\sigma}_j}, \qquad q = s_{(\lceil (m+1)(1-\alpha) \rceil)}, \tag{12}$$

where j ranges over the pooled (profile, time) pairs of the calibration set, m is their number, and $s_{(k)}$ denotes the k-th order statistic. Setting $\kappa = q$ yields a band whose coverage on exchangeable test profiles is at least $1 - \alpha$ by the standard split-conformal guarantee [37]; no distributional assumption on the fused errors is involved. The fused variance $\hat{\sigma}^2(t)$ acts as the score's scale function: a well-calibrated $\hat{\sigma}$ yields $q \approx z_{\alpha/2}$, whereas a miscalibrated one is corrected through q at the price of a wider or narrower band. Coverage being guaranteed for every method under this construction, the discriminating quality metric across methods becomes the width of the calibrated band. The remaining storage life is the first time T_f at which $\hat{P}(t)$ crosses the failure threshold P_{fail}. Its prediction interval comes from the upper and lower bounds of the calibrated band, read off where they cross P_{fail}. The upper bound of the band crosses later and gives the upper end of the T_f interval. The lower bound crosses earlier and gives the lower end. The construction generalizes to a degradation parameter that increases toward an upper failure threshold by reversing the threshold comparison.

For a degradation parameter that decreases with time (the stress-relaxation case),

$$\hat{T}_f = \inf\{\, t : \hat{P}(t) \le P_{\text{fail}} \,\}, \tag{13}$$

$$T_f^{\text{lo}} = \inf\{\, t : \hat{P}(t) + \kappa\,\hat{\sigma}(t) \le P_{\text{fail}} \,\}, \tag{14}$$

$$T_f^{\text{hi}} = \inf\{\, t : \hat{P}(t) - \kappa\,\hat{\sigma}(t) \le P_{\text{fail}} \,\}, \tag{15}$$

where $\kappa = q$ from Equation (12) when a calibration population is available, and $\kappa = z_{\alpha/2}$ as the uncalibrated Gaussian reference on a single real trajectory. The reported $1 - \alpha$ prediction interval is $[T_f^{\text{lo}}, T_f^{\text{hi}}]$; when the upper band does not cross P_{fail} within the prediction horizon, T_f^{hi} is right-censored at the horizon and the interval is reported as unbounded above. For a degradation parameter that increases toward an upper failure threshold, the same construction applies with the threshold comparison reversed.

The interval inherits two distinct geometric contributions. The predictive variance $\hat{\sigma}^2(t)$ at the crossing time sets the vertical width of the band; the slope of $\hat{P}(t)$ at the crossing converts that vertical width to a horizontal time interval. A steep slope at the crossing produces a narrow T_f interval even when the band is wide; a shallow slope produces a wide T_f interval even when the band is narrow. The construction thus reflects two distinct sources of difficulty in pinpointing T_f: how uncertain the trajectory is, and how slowly it approaches the threshold.

Calibration is established and verified on a Monte Carlo population with known truths. Each run draws its own ground-truth degradation law including an independently drawn secondary-stage slope, generates accelerated and natural data with measurement noise, and runs the full pipeline; the population thereby contains profile-to-profile variation in

the truth itself, not merely noise realizations around one fixed trajectory. The runs are split into disjoint calibration and test halves: the conformal quantile q of Equation (12) is computed once on the calibration half and applied unchanged to the test half, so that every reported coverage number is out-of-sample by construction. Empirical first-passage coverage is the fraction of test runs in which the true T_f falls inside $[T_f^{\text{lo}}, T_f^{\text{hi}}]$, with the censoring convention stated above. A complementary trajectory-level pointwise coverage,

$$\text{Cov}_{\text{traj}} = \frac{1}{N}\sum_{r=1}^{N}\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\mathbf{1}\Big\{P_{\text{truth}}^{(r)}(t) \in \big[\hat{P}^{(r)}(t) \pm q\,\hat{\sigma}^{(r)}(t)\big]\Big\}, \tag{16}$$

checks pointwise calibration of the calibrated band across the full prediction grid $\mathcal{T}$ over the N test runs, with $P_{\text{truth}}^{(r)}$ the run-specific truth. The two metrics together test calibration at the first-passage level and at the trajectory level. Section 3 reports both, alongside the mean calibrated band width as the sharpness metric that discriminates the methods once coverage is guaranteed.

2.4. Algorithm and Implementation

Algorithm 1 consolidates the procedure from end to end. Each line maps to a specific equation in Sections 2.2 and 2.3.

Algorithm 1: Empirically calibrated multi-fidelity fusion with conformal T_f interval

Input: natural-storage data $\mathbf{Y}^{(n)}$, accelerated-storage data $\{\mathbf{Y}^{(a,k)}\}_{k=1}^{K}$, initial value P_0, failure threshold P_{fail}, confidence level $1-\alpha$, prediction grid $\mathcal{T}$, optional calibration population for the conformal quantile q

Output: fused trajectory $\hat{P}(t)$, predictive variance $\hat{\sigma}^2(t)$, point estimate $\hat{T}_f$, prediction interval $[T_f^{\text{lo}}, T_f^{\text{hi}}]$, diagnostic $\bar{\rho}$

`// Per-temperature regression (accelerated branch)`

for $k = 1, \ldots, K$ **do**
  Fit Equation (3) on $\mathbf{Y}^{(a,k)}$; obtain $(\hat{A}_k, \hat{B}_k, \hat{\sigma}_k^2)$ with standard errors;
end

`// Arrhenius extrapolation to operating temperature`

WLS Equation (4) on $\ln \hat{B}_k$ vs. $1/T_k$ gives $\hat{B}_{\text{op}}$ and $\text{Var}(\hat{B}_{\text{op}})$;

Analogous WLS gives $\hat{A}_{\text{op}}$ and $\text{Var}(\hat{A}_{\text{op}})$; pool $\hat{\sigma}_k^2$ to $\sigma^2_{\text{intrinsic}}$;

`// Additive empirical model-form variance`

Compute training residuals $r(t_i) = Y_i^{(n)} - \hat{P}_a(t_i)$ from Equation (1);

Fit Equation (6) on $\{(t_i, r^2(t_i))\}$ with $a, b \geq 0$; set $\sigma^2_{\text{MF}}(t)$ and $\sigma_a^2(t)$ via Equation (7); record $\bar{\rho}$ via Equation (8);

`// Natural-branch model selection and prediction`

Fit $\mathcal{M}_1$ and $\mathcal{M}_2$ of Equation (9) on $\mathbf{Y}_{\text{train}}^{(n)}$; select by AICc; obtain $(\hat{P}_n(t), \hat{\sigma}_n^2(t))$ on $\mathcal{T}$ from the selected design;

`// Precision-weighted fusion`

for $t \in \mathcal{T}$ **do**
  Compute $\alpha_n(t)$ via Equation (10); no caps or thresholds apply;
  Compute $\hat{P}(t)$ and $\hat{\sigma}^2(t)$ via Equation (11);
end

`// Conformal calibration and first-passage interval`

If a calibration population is available, compute q via Equation (12) and set $\kappa = q$; otherwise, set $\kappa = z_{\alpha/2}$ as the uncalibrated reference;

Locate $\hat{T}_f$, T_f^{lo}, T_f^{hi} on $\mathcal{T}$ via Equations (13)–(15) by linear interpolation between adjacent grid points;

return $\hat{P}(t)$, $\hat{\sigma}^2(t)$, $\hat{T}_f$, $[T_f^{\text{lo}}, T_f^{\text{hi}}]$, $\bar{\rho}$;

All steps are closed-form OLS or WLS with delta-method propagation; the AICc selection compares two closed-form fits, and the conformal step is a single order statistic. The procedure contains no MCMC, no kernel hyperparameter optimization, no tuned constants, and no iterative inner loop other than the prediction-grid evaluation. Wall-clock cost is dominated by the per-temperature regressions and scales linearly in the total

accelerated-storage data size. For the two case studies in Section 3, a single complete fit runs in under 0.1 s on a standard workstation. The lower and upper endpoints T_f^{lo} (Equation (14)) and T_f^{hi} (Equation (15)) of the prediction interval together quantify the operational uncertainty on the remaining storage life.

3. Results

This section reports the experimental validation of the proposed method against five baselines on two real case studies and a synthetic Monte Carlo simulation. Section 3.1 describes the datasets, baselines, evaluation metrics, and implementation. Section 3.2 validates the multi-fidelity fusion with decomposed variance and empirical calibration on the torsion-bar and amplifier case studies. Section 3.3 validates the conformal prediction interval on a varying-truth Monte Carlo population with disjoint calibration and test halves. Section 3.4 names the cross-case patterns. Section 3.5 reports sensitivity to the held-out window length, a low-signal-to-noise variant of the amplifier case, and four distributional diagnostics requested by the analysis: bootstrap stability of the model-form variance fit, residual normality tests, a sampling-based check of the delta-method propagation, and a sampled first-passage-time distribution.

3.1. Experimental Setup

Two case studies represent two distinct degradation regimes in aerospace dormant components. The torsion bar is a mechanical-storage case with 32 specimens monitored over 120 months of natural storage; the degradation parameter is the holding torque, with $P_0 = 35.030$ N·m and failure declared at $P_{\mathrm{fail}} = 28$ N·m. Accelerated-storage data come from 8 specimens per temperature at four temperatures $\{110, 130, 150, 170\}$ °C over 144 h. The amplifier (Rudder III) is an electronic-storage case with 12 specimens monitored over 13 years; the degradation parameter is the absolute output current, with $|I_0| = 10.013$ mA and failure declared at $|I - I_{\mathrm{nominal}}| > 0.15$ mA. Accelerated-storage data come from 3 specimens per temperature at $\{90, 100, 110, 120\}$ °C over 130–250 days, depending on the temperature.

The held-out protocol withholds the most recent natural-storage observations to assess extrapolation accuracy. For the torsion bar, the last 24 months (20% of the storage horizon) are held out. For the amplifier, the last 3 years (23% of the horizon) are held out. The fusion procedure consumes only training-window natural-storage data plus the full accelerated-storage set, and performance is evaluated against the held-out natural-storage observations.

We compare against five baselines that span the relevant fusion strategies: Natural-only fits an OLS regression on training-window natural data, ignoring accelerated data; it uses the same AICc-selected mechanism-specific model as the proposed method's natural branch, so that the comparison isolates the fusion machinery rather than the choice of mean model. Accelerated-only uses Arrhenius extrapolation to the operating temperature, ignoring natural data. Naive average fuses with a fixed scalar weight $\alpha_n = 0.5$. KL-entropy weighting [13] computes a scalar weight from the relative-entropy difference between source predictions on the training grid; this is the method developed in a prior Chinese aerospace-reliability line of work and is the most direct fusion competitor. Bayesian calibration factor [14] fits a posterior on a scalar K that calibrates the accelerated prediction against natural training data and is the canonical Bayesian fusion baseline. Implementations follow the same data conventions as the proposed method to ensure a fair comparison.

Four metric groups capture point accuracy, interval validity, interval sharpness, and time-varying behavior. The held-out root-mean-square error (RMSE) quantifies point accuracy on the masked natural-storage window. On the real cases, where only a single trajectory exists, the held-out 95% coverage of the uncalibrated Gaussian reference band

($\kappa = 1.96$) is reported with the explicit caveat that 24 and 3 held-out points carry little inferential weight. On the simulation track, calibration is established by the split-conformal construction of Section 2.3: the conformal quantile q, the out-of-sample trajectory coverage of Equation (16), the mean calibrated band width as the sharpness metric, and the first-passage coverage of the T_f interval under the right-censoring convention. The natural-branch weight $\alpha_n(t)$ at fixed times reports the fusion mixture.

All methods are implemented in Python 3.11 using NumPy, pandas, and SciPy; no MCMC, kernel hyperparameter optimization, or gradient-based learning is required. The proposed method's Arrhenius regression and empirical-calibration steps are closed-form, as is the precision-weighted fusion at each prediction time. Each complete fit runs in under 0.1 s on a standard workstation. The simulation in Section 3.3 runs a 1000-run varying-truth population (500 calibration + 500 disjoint test runs) and a 400-run fixed-truth control (200 + 200), all with seed 20260611. The runtime is approximately one minute from end to end.

3.2. Validation of the Decomposed-Variance Fusion

This subsection tests two hypotheses about the fusion method's behavior on real data: First, the proposed fusion remains close to the natural-only estimator on point accuracy regardless of how misspecified the accelerated branch is, where every existing fusion method is pulled away by Arrhenius extrapolation error. Second, the proposed fusion assigns natural and accelerated branches weights that track each branch's instantaneous trustworthiness, where existing methods assign a single scalar weight independent of time. The two hypotheses together test whether the method delivers safe-fallback fusion. The target behavior is at least as informative as natural-only on point estimates, with a continuous data-driven weight that diagnoses the accelerated branch, and an interval that survives population-level calibration without ad hoc width.

Figure 2 displays the decomposition of $\sigma_a^2(t)$ from Equation (5) on the torsion data, in normalized loss units squared. The intrinsic component, pooled across the four accelerated temperatures, is approximately constant in time at $\sigma^2_{\text{intrinsic}} \approx 4.47 \times 10^{-5}$. The OLS-propagated extrapolation component grows with time through the $\varphi^2(t)$ factor, rising from $\approx 5 \times 10^{-7}$ at $t = 1$ month to 3.25×10^{-6} at $t = 20$ years. The additive model-form variance $\sigma^2_{\text{MF}}(t)$ of Equation (7), fitted from the natural-training residuals, rises from 5.5×10^{-3} to 3.8×10^{-2} over the same window: three to four orders of magnitude above the OLS term and the dominant contribution to $\sigma_a^2(t)$ throughout. The gap between the OLS-propagated and the model-form components is the quantitative signature of Arrhenius misspecification at room temperature.

The model-form variance of Equation (6) is fitted on the residuals between accelerated prediction and natural-storage training observations. On the torsion case, the model-form variance ratio of Equation (8) averages $\bar{\rho} \approx 11{,}600$ over the prediction horizon (bootstrap 95% interval 9700–14,600; Section 3.5), signaling severe Arrhenius model-form mismatch when extrapolating from the 110–170 °C test window down to room temperature. As a consequence, the precision-weighted fusion of Equation (10) assigns $\alpha_n(t) \approx 0.998$ throughout the prediction window, with the accelerated branch contributing the remaining $\approx$0.002 as a residual prior on the long-term trend. The same diagnostic on the amplifier case yields $\bar{\rho} \approx 81$ and $\alpha_n(t)$ between 0.95 and 0.96 across the prediction window. No threshold or cap is involved; the weights are the continuous output of Equations (7) and (10).

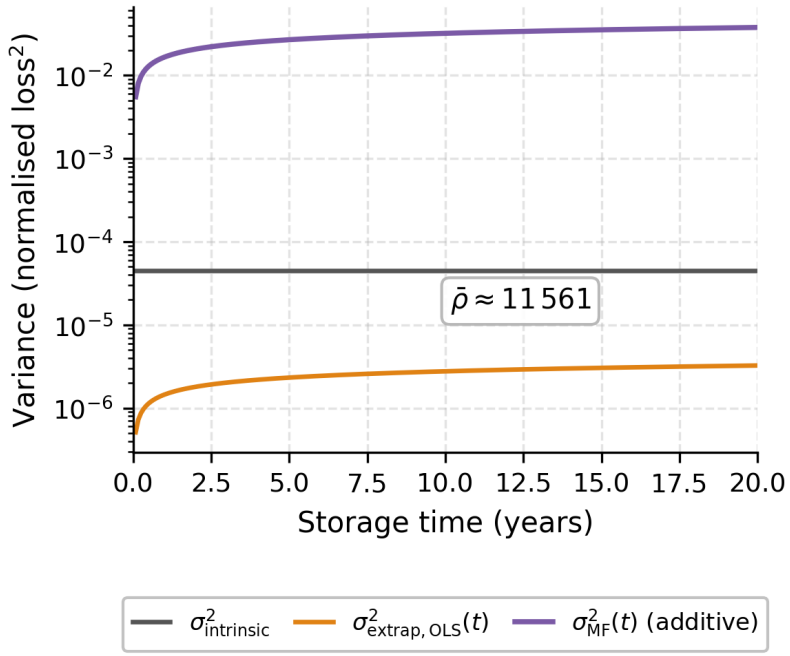
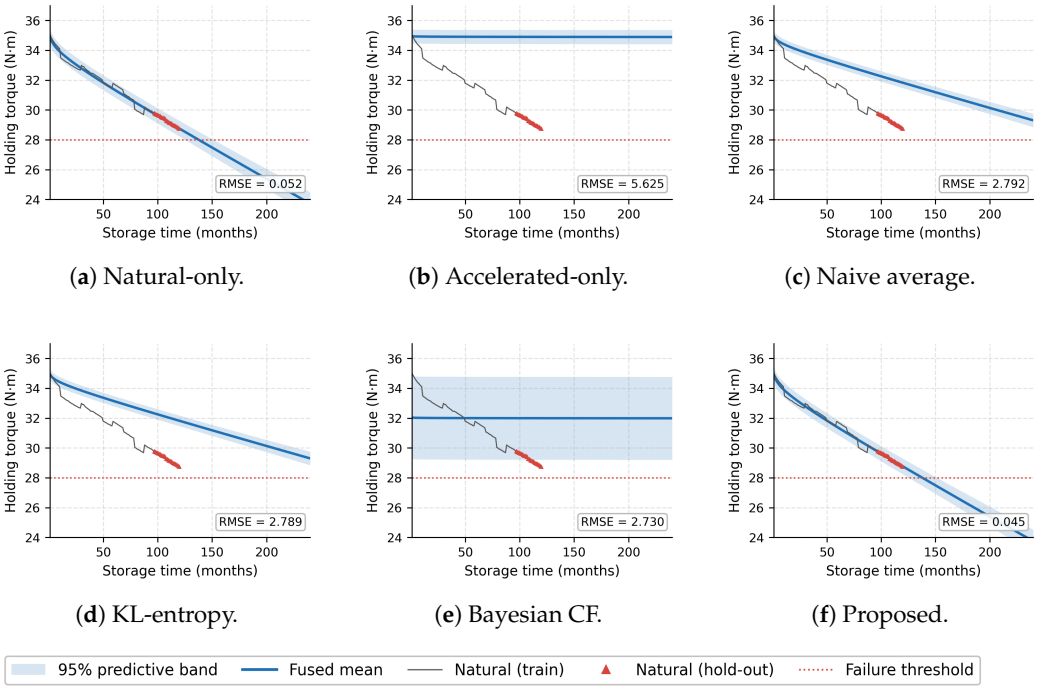


Figure 2. Variance decomposition of the accelerated-storage branch on the torsion case. Values in normalized-loss-squared units on a logarithmic axis. The intrinsic component $\sigma^2_{\text{intrinsic}}$ is pooled across the four accelerated temperatures and is essentially constant in time. The OLS-propagated sampling component $\sigma^2_{\text{extrap,OLS}}(t)$ grows with $\varphi^2(t)$ but stays an order of magnitude below the intrinsic level over the 20-year horizon. The additive model-form variance $\sigma^2_{\text{MF}}(t)$ of Equation (7), fitted from natural-storage residuals, sits three to four orders of magnitude above the OLS term, giving the model-form ratio $\bar{\rho} \approx 11{,}600$ of Equation (8). This gap is the quantitative signature of Arrhenius misspecification at room temperature and drives $\alpha_n(t) \to 1$ on this case.

The held-out trajectories in Figure 3 show the consequence on point prediction. On this case, the AICc selection of Equation (9) decisively prefers the two-member relaxation form: AICc is -261.6 for $\mathcal{M}_2$ ($\ln t$ plus secondary linear term) against -109.9 for $\mathcal{M}_1$ ($\ln t$ only), a difference of about 152. The fitted secondary slope corresponds to 0.48 N·m of additional torque loss per year, a contribution that is invisible within the 144 h accelerated window but dominant at decade horizons. The natural-only and proposed-method panels track the held-out observations closely. The three fusion baselines (naive-average, KL-entropy, Bayesian calibration factor) deviate upward, pulled by the biased accelerated prediction. The accelerated-only panel remains near the initial value and fails to capture the magnitude of natural-storage degradation. The held-out RMSE values are summarized in Table 1: 0.045 N·m for the proposed method, 0.052 N·m for natural-only, and 2.730 N·m to 5.625 N·m for the four other baselines. The proposed method slightly improves on natural-only (by about 13%) and improves over the next-best fusion baseline (Bayesian calibration factor) by a factor of 61; the residual bias of the fused mean on the held-out window is -0.03 N·m.

The amplifier case in Figure 4 shows the same pattern with the orientation reversed: the degradation parameter increases with time toward an upper failure threshold. Here, the AICc selection retains the primary-only drift form (AICc -93.4 for linear against -91.9 for the two-member form): no early logarithmic transient is detectable in the natural stream, and the selector correctly declines the extra term. The held-out RMSE values are 11.4 µA for the proposed method, 9.6 µA for natural-only, and 32.1 µA to 56.0 µA for the four other baselines. The proposed method stays within 1.8 µA (19%) of natural-only, while

the three scalar-weight fusion baselines have RMSE values 2.8 to 3.1 times larger than the proposed method's.

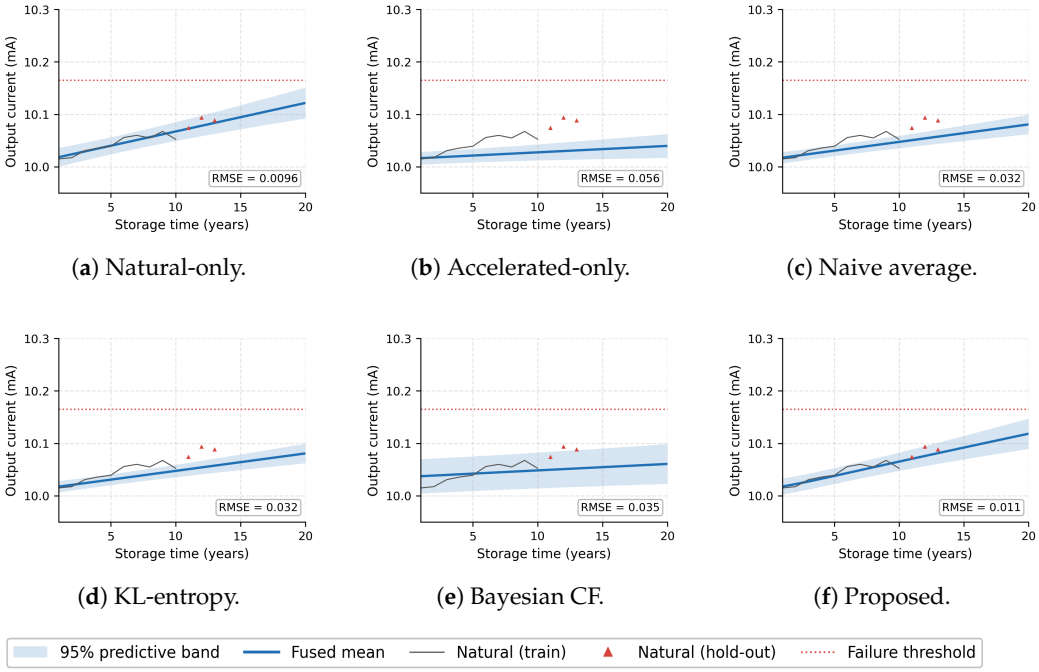


(**a**) Natural-only. (**b**) Accelerated-only. (**c**) Naive average.

(**d**) KL-entropy. (**e**) Bayesian CF. (**f**) Proposed.

Figure 3. Fused trajectory of holding torque on the torsion case under six fusion strategies. The last 24 months of natural-storage data are held out for evaluation. Each panel shows the monthly natural-storage training means as a thin gray line, the held-out means as red triangles, the fused mean and its 95% Gaussian-reference band for that method, and the failure threshold $P_{\text{fail}} = 28$ N·m. With the AICc-selected natural model, the proposed method in panel (**f**) and natural-only in panel (**a**) track the held-out trajectory closely (RMSE 0.045 and 0.052 N·m); the three scalar-weight fusion baselines in panels (**c**–**e**) are pulled upward by the biased accelerated branch, and the accelerated-only panel (**b**) misses the degradation magnitude entirely. The held-out RMSE for each method is annotated in the lower-right corner of its panel.

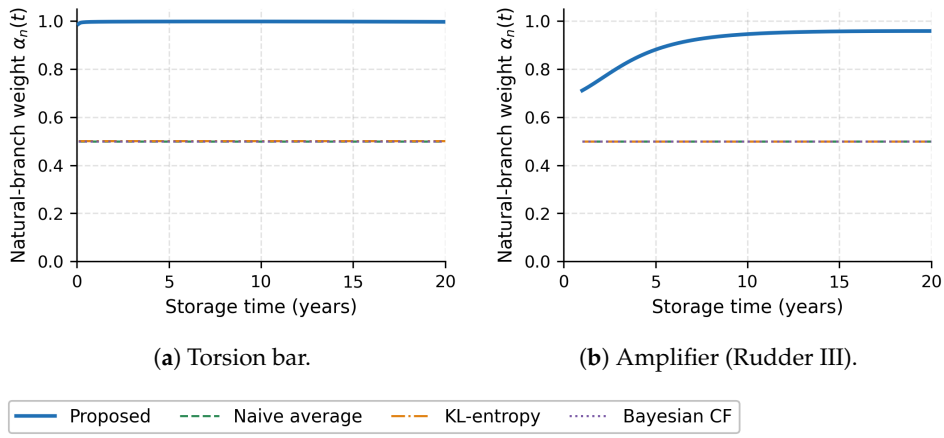
Figure 5 reveals the structural mechanism behind these results. The proposed method's $\alpha_n(t)$ sits at the top of both panels (about 0.998 on torsion, between 0.95 and 0.96 on the amplifier), while the three scalar-weight baselines lock at $\alpha_n \approx 0.5$ as their formulation requires; the proposed weights arise continuously from Equation (10), with no threshold anywhere. Both case studies thus confirm the hypotheses. Under the diagnosed strong Arrhenius misspecification, the proposed method's time-varying weight shifts toward unity and recovers the natural-only point estimate. The scalar-weight fusion baselines cannot make this shift and remain pulled by the biased accelerated prediction. The fusion's value on these cases is therefore threefold: it matches or slightly improves on natural-only in point accuracy, it detects the accelerated-branch bias from the data through $\bar{\rho}$, and it carries a predictive variance that survives population-level calibration without ad hoc inflation (Section 3.3).



(**a**) Natural-only. (**b**) Accelerated-only. (**c**) Naive average.

(**d**) KL-entropy. (**e**) Bayesian CF. (**f**) Proposed.

Figure 4. Fused trajectory of $|I|$ on the amplifier Rudder III case under six fusion strategies. The last 3 years of natural-storage data are held out for evaluation. Each panel shows the yearly natural-storage means as a thin gray line (training) and red triangles (held-out), the fused mean and its 95% Gaussian-reference band, and the failure threshold $|I| = 10.165$ mA as a dotted red line. The AICc selection retains the primary-only linear-drift form on this case. The proposed method in panel (**f**) and natural-only in panel (**a**) are the only strategies whose mean approaches the held-out triangles (RMSE 11.4 and 9.6 µA); the four other baselines under-predict drift across the held-out window.



(**a**) Torsion bar. (**b**) Amplifier (Rudder III).

Figure 5. Time-varying natural-branch fusion weight $\alpha_n(t)$ on the two case studies, for the proposed method and the three scalar-weight fusion baselines. The scalar baselines (naive average, KL-entropy, Bayesian calibration factor) lock at $\alpha_n = 0.5$ regardless of time, reflecting their formulation. The proposed method's weight (thick blue line) sits near unity throughout both prediction windows—about 0.998 on the torsion case, and between 0.95 and 0.96 on the amplifier—as the continuous output of Equation (10) with the additive model-form variance of Equation (7); no threshold, cap, or switching rule is involved. This continuous behavior is the structural feature that lets the proposed method fall back safely to the natural-only estimator on real cases.

Table 1. Held-out RMSE and Gaussian-reference 95% coverage on the two real case studies. Coverage on a single real trajectory uses the uncalibrated Gaussian reference band ($\kappa = 1.96$) and a small held-out sample (24 and 3 points, respectively); population-level calibration is established on the simulation of Table 2. The best per column (excluding accelerated-only, the unfused Arrhenius reference) is highlighted in bold. With the AICc-selected natural model, the proposed method matches or slightly improves on natural-only in point accuracy, while every scalar-weight fusion baseline is pulled away by the biased accelerated branch.

Method	Torsion Bar		Amplifier Rudder III	
	RMSE (N m)	Cov. (%)	RMSE (μA)	Cov. (%)
Natural-only	0.052	**100.0**	**9.6**	**100.0**
Accelerated-only	5.625	0.0	56.0	0.0
Naive-average	2.792	0.0	32.1	0.0
KL-entropy	2.789	0.0	32.1	0.0
Wang-2013-BCF	2.730	54.2	35.5	33.3
Proposed	**0.045**	**100.0**	11.4	**100.0**

3.3. Validation of the Conformal Prediction Interval

This subsection tests two properties of the proposed method's 95% prediction interval on a Monte Carlo population with known truths: validity, meaning out-of-sample coverage at the nominal level on profiles disjoint from the calibration set; and sharpness, meaning the narrowest calibrated band among the six methods. Trajectory-level pointwise coverage, first-passage coverage of the T_f interval, and the mean calibrated band width are reported.

The primary population is a varying-truth design: each run draws its own ground-truth stress-relaxation law from Equation (1), extended by a secondary linear stage, with $E_a \sim \mathcal{U}(0.18, 0.30)$ eV, $\ln B_0 \sim \mathcal{N}(\ln 133, 0.25^2)$, $A_0 \sim \mathcal{U}(0.002, 0.010)$, and a room-temperature secondary slope $C \sim \mathcal{U}(0, 4 \times 10^{-5})$ per day, Arrhenius-scaled to the test temperatures. The secondary term contributes at or below the accelerated noise floor within the 144 h test window but up to 0.29 in normalized loss at 20 years, so the accelerated branch is misspecified by construction in a way that mirrors the real cases; the median model-form ratio across runs is $\bar{\rho} \approx 2700$, against ≈ 1 in a fixed-truth control population that reuses the submitted design (single truth, $E_a = 0.22$ eV, $T_f = 19.74$ years, $C = 0$). Each run generates accelerated data (four temperatures, eight specimens, 13 time points each, noise SD 0.005) and a 120-month natural stream (monthly-mean noise SD $0.005 P_0/\sqrt{32}$), runs all six methods end to end, and records the predictive band on the 240-month grid together with the band-implied T_f interval. The 1000 varying-truth runs are split into 500 calibration and 500 disjoint test runs; the control uses 200 plus 200. The conformal quantile q of Equation (12) is computed once per method on the calibration half and applied unchanged to the test half.

Figure 6 and Table 2 report the test-half results on the varying-truth population. Out-of-sample trajectory coverage lands between 94.5% and 95.6% for all six methods, as the split-conformal guarantee requires; validity therefore no longer discriminates the methods, and the discriminating metric is the width each method needs to attain it. The proposed method reaches the guaranteed coverage at a mean band width of 0.071 N·m, essentially tied with natural-only (0.072 N·m) and 97 to 196 times narrower than the four other fusion baselines (6.9 N·m to 14.0 N·m). The conformal quantile itself diagnoses each method's nominal variance: the proposed method needs $q = 1.01$, meaning that its fused $\hat{\sigma}(t)$ is already population-calibrated and the conformal step changes almost nothing, whereas accelerated-only needs $q = 38.7$ to stretch a wildly over-confident band to validity. The AICc selector is validated in passing: it picks the two-member relaxation form in 99.1% of the varying-truth

runs, whose truths almost certainly contain a secondary stage, and the primary-only form in 85.8% of the fixed-truth control runs, whose truth is exactly logarithmic.

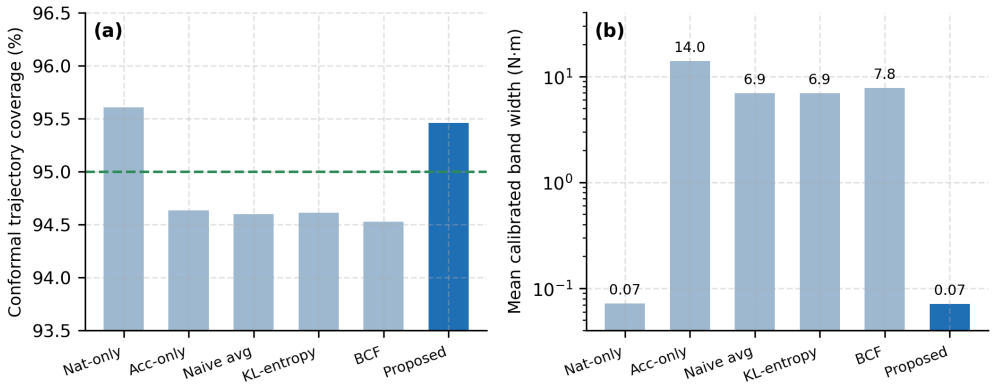


Figure 6. Split-conformal calibration on the varying-truth Monte Carlo population (500 calibration + 500 disjoint test runs). Panel (**a**): Out-of-sample trajectory coverage of the conformally calibrated band on the test half. All six methods land between 94.5% and 95.6%, as the split-conformal guarantee requires; the dashed green line marks the nominal 95%. Panel (**b**): The price each method pays for that validity, measured as the mean calibrated band width on a logarithmic axis. The proposed method attains the guaranteed coverage at 0.071 N·m, essentially tied with natural-only and 97 to 196 times narrower than the four other fusion baselines. Once coverage is guaranteed by construction, sharpness is the discriminating metric, and the proposed method dominates it among the fusion strategies.

Table 2. Population-level Monte Carlo with split-conformal calibration. Primary varying-truth population: 500 calibration + 500 disjoint test runs, each with an independently drawn ground truth including a random secondary-stage slope. The conformal quantile q (Equation (12)) is computed on the calibration half only; all reported numbers are out-of-sample on the test half. Coverage of the calibrated band is guaranteed near the nominal 95% for every method; the discriminating metric is the mean calibrated band width (sharpness). q itself diagnoses each method's variance calibration: $q \approx 1$ means the fused $\hat{\sigma}$ is already population-calibrated, $q \gg 1.96$ means the nominal variance is badly over-confident. T_f coverage uses the right-censoring convention of Section 2.3; 30.4% of the proposed-method test intervals are upper-censored at the horizon.

Method	q	Traj. Cov. (%)	Band Width (N m)	T_f Cov. (%)	RMSE-Truth
Natural-only	1.02	95.6	0.072	97.5	0.0148
Accelerated-only	38.72	94.6	13.99	100.0	2.929
Naive-average	37.68	94.6	6.94	100.0	1.465
KL-entropy	37.68	94.6	6.92	100.0	1.462
Wang-2013-BCF	6.17	94.5	7.81	100.0	2.186
Proposed	1.01	95.5	**0.071**	96.9	**0.0148**

Bold entries mark the best (smallest) value in the column among all six methods.

First-passage coverage on the varying-truth test half is 96.9% for the proposed method, computed over the 358 test runs whose truth crosses the threshold within 600 months (median truth $T_f = 162$ months) and using the right-censoring convention of Section 2.3; 30.4% of the proposed intervals are upper-censored at the 240-month horizon. The fixed-truth control retains, by construction, the window-edge artifact of the submitted version; its truth $T_f = 19.74$ years sits within 0.3 years of the horizon, 93% of intervals are upper-censored, and first-passage coverage is then driven mostly by the censoring rule. The varying-truth population, whose failure times sit well inside the horizon, is therefore the primary first-passage testbed, and on it the metric confirms calibration.

The behavior on real data appears in Figure 7. The proposed method's mean (solid curve) tracks both the training-window natural-storage observations and the held-out

triangles, and all 24 held-out observations fall inside the uncalibrated Gaussian reference band ($\kappa = 1.96$), whose half-width is approximately 0.52 N·m across the held-out window. With the corrected natural model, the fused mean now crosses the 28 N·m failure threshold inside the horizon: the point estimate is $\hat{T}_f = 137.9$ months (11.5 years), and the band-crossing interval of Equations (14) and (15) is $(125.4, 151.4)$ months—that is, $(10.4, 12.6)$ years. A sampled-trajectory cross-check (Section 3.5) gives a parameter-uncertainty first-passage interval of $(132.8, 143.9)$ months, strictly inside the geometric interval, so the geometric construction errs on the conservative, operationally safe side. Coverage on a single real trajectory carries little inferential weight; the calibration claim rests on the population-level conformal validation above.

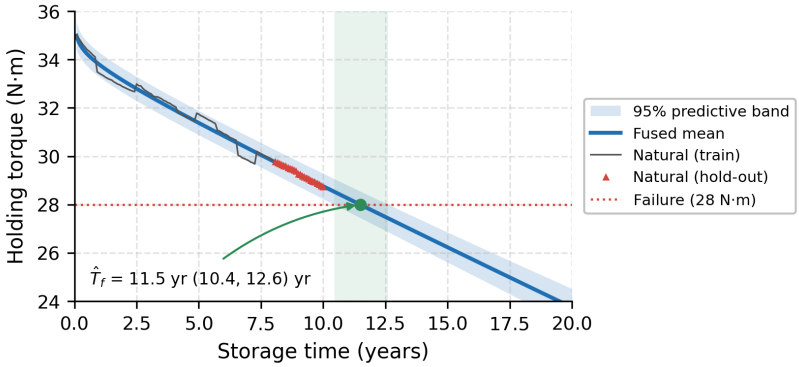


Figure 7. Proposed-method fused trajectory and 95% Gaussian-reference band on the torsion case. The fused mean (solid blue) tracks the natural-storage training observations (thin gray line) and all 24 held-out observations (red triangles), each of which falls inside the band; the band half-width is approximately 0.52 N·m across the held-out window. With the AICc-selected natural model, the mean crosses the failure threshold $P_{\text{fail}} = 28$ N·m inside the horizon at $\hat{T}_f = 11.5$ years (green marker), and the band-crossing construction of Equations (14) and (15) gives the remaining-storage-life interval $(10.4, 12.6)$ years (green shading). A sampled-trajectory cross-check places the parameter-uncertainty first-passage interval at $(11.1, 12.0)$ years, strictly inside the geometric interval, so the reported interval is conservative in the operationally safe direction.

Under conformal calibration, every method's band is valid by construction, so the engineering question shifts to what that validity costs. The proposed method pays the least: it attains the guaranteed 95% at a band two orders of magnitude narrower than any other fusion strategy, with a nominal variance so close to calibrated that the conformal correction is nearly the identity. It is the only fusion method whose interval is simultaneously valid and sharp enough to support storage-life decisions.

3.4. Joint Cross-Case Patterns

Two cross-case patterns emerge across the two real cases and the simulation. The first pattern concerns what fusion delivers when the accelerated branch is biased. On both real cases, the accelerated branch is diagnosed as strongly misspecified by the model-form variance. The proposed fusion matches or slightly improves on the natural-only baseline in terms of point accuracy. The four baselines that lack the model-form term are pulled away from the natural-only trajectory by the biased accelerated branch, and their held-out RMSE values range from 2.8 to 126 times the proposed method's across the two cases. The contribution of the proposed fusion on these cases is threefold: First, the mechanism-specific AICc selection removes the natural-branch form error that any fusion would otherwise inherit. Second, the method detects accelerated-branch bias from data through $\bar{\rho}$ and falls back continuously, where existing fusion methods cannot. Third, the method's predictive

variance is the only fused variance that passes population-level conformal calibration at a usable width.

The second pattern concerns the role of the variance decomposition. The decomposition of $\sigma_a^2(t)$ into intrinsic, sampling, and model-form components lets the additive σ_{MF}^2 register what the OLS-propagated variance cannot. The model-form ratio spans four orders of magnitude across the test set: $\bar{\rho} \approx 11{,}600$ on the torsion bar, ≈81 and ≈61 on the two amplifier variants, ≈2700 in the varying-truth simulation whose truths contain a hidden secondary stage, and ≈1 in the fixed-truth control where the Arrhenius model is exact by construction. The time-varying weight $\alpha_n(t)$ adjusts continuously across this entire span, with no thresholds. In the fixed-truth control, where nothing is misspecified, the proposed method relaxes toward the precision-weighted balance and remains calibrated; in every misspecified regime, it converges to the natural-only estimator. The variance decomposition therefore plays two roles: it diagnoses each case from training data alone, and it provides the mechanism that recovers idealized fusion behavior exactly when the model assumptions hold.

3.5. Sensitivity and Limits

The safe-fallback claim is robust to the choice of held-out window length, and the longest window exposes an honest limit. Figure 8 reports the held-out RMSE for the torsion case under four held-out window lengths: 12, 24, 36, and 48 months. The natural-only baseline grows from 0.023 N·m at 12 months to 0.683 N·m at 48 months; the proposed method tracks it within 0.01 N·m at every window, growing from 0.030 N·m to 0.691 N·m. The three intermediate fusion baselines stay flat between 2.7 N·m and 3.0 N·m, and the accelerated-only baseline stays above 5.2 N·m. The Gaussian-reference coverage of the proposed band over the same windows is 100%, 100%, 91.7%, and 12.5%; with only 72 training months, the secondary-stage slope is estimated from a shorter window and under-predicts the curvature of the final four years, and the band, whose width is honest rather than inflated, misses accordingly. The single-trajectory extrapolation horizon of the present formulation is therefore in the order of one third of the training window, which is precisely why interval validity is claimed at the population level through conformal calibration rather than from any single trajectory.

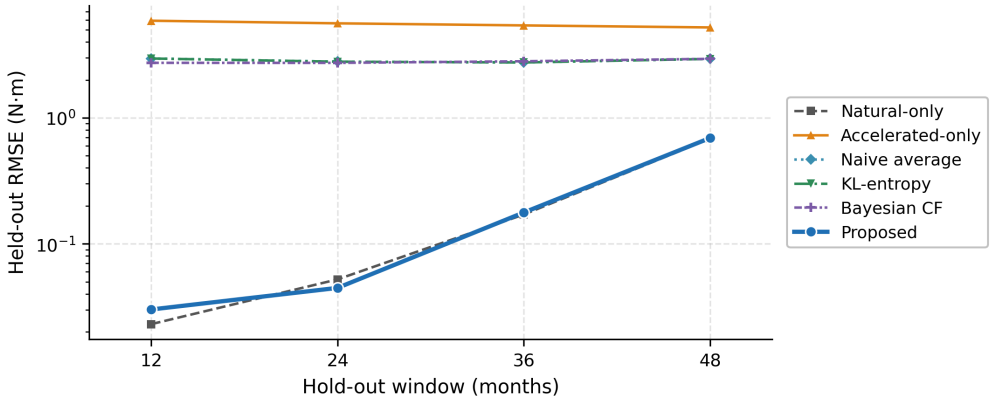
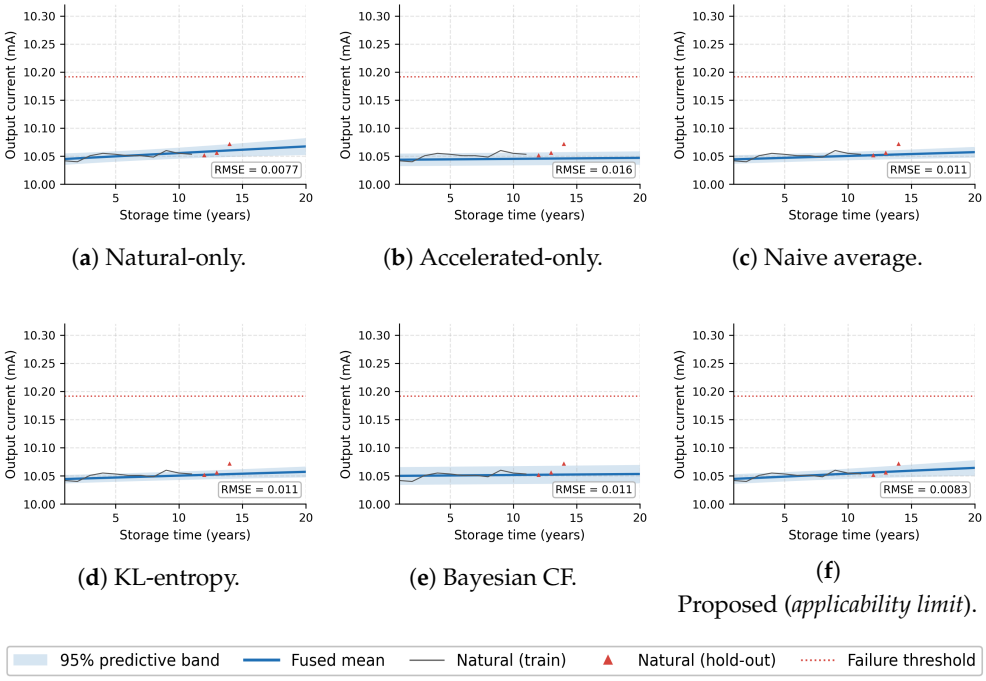


Figure 8. Held-out RMSE on the torsion case as a function of the held-out window length, for windows of 12, 24, 36, and 48 months (logarithmic axis). The proposed method (thick blue) and the natural-only baseline (gray) overlap near the bottom of the panel and stay within 0.01 N·m of each other at every window, rising together from about 0.03 N·m at the 12-month window to about 0.69 N·m at the 48-month window as the training window shrinks and the secondary-stage slope becomes harder to identify. The three scalar-weight fusion baselines stay flat between 2.7 and 3.0 N·m, and the accelerated-only baseline above 5.2 N·m, both dominated by the accelerated-branch bias rather than by the window choice.

The safe fallback to natural-only is stress-tested where natural-only is itself near the noise floor. Figure 9 reports the low-signal-to-noise variant of the amplifier case: Rudder II, whose natural-storage drift is comparable in magnitude to the year-to-year measurement noise. The held-out RMSE values are 7.7 µA for natural-only, 8.3 µA for the proposed method, 10.9 µA to 11.4 µA for the scalar-weight fusion baselines, and 16.4 µA for accelerated-only. The proposed method's gap to natural-only is 7%; the model-form variance ($\bar{\rho} \approx 61$) keeps $\alpha_n(t) \approx 0.85$ across the horizon through the continuous weighting alone, with no cap or threshold anywhere in the pipeline. The residual 7% is the price of retaining a nonzero accelerated weight in a regime where that branch is diagnosed as unreliable and the natural branch is itself noisy. Gaussian-reference coverage on the three held-out points is 66.7% (2 of 3) for the proposed method and all fusion baselines, and 100% for natural-only; with $n = 3$, these proportions carry essentially no inferential weight. The case characterizes the applicability boundary of the method: when $\bar{\rho} \gg 1$ coincides with a low natural-branch signal-to-noise ratio, fusion cannot beat natural-only, and the method's value reduces to flagging exactly that condition from training data.



(**a**) Natural-only. (**b**) Accelerated-only. (**c**) Naive average.

(**d**) KL-entropy. (**e**) Bayesian CF. (**f**) Proposed (*applicability limit*).

Figure 9. Same six-method comparison as Figure 4, applied to the low-signal-to-noise variant Rudder II. The yearly means (thin gray line and red triangles) show year-to-year scatter comparable in magnitude to the underlying drift, contrasting with the cleaner Rudder III signal. The natural-only baseline in panel (**a**) retains a 7% edge over the proposed method in panel (**f**) (RMSE 7.7 versus 8.3 µA); the model-form variance keeps the continuous weight at $\alpha_n(t) \approx 0.85$ with no cap anywhere in the pipeline, and the residual gap is the price of the nonzero accelerated weight in a regime where that branch is diagnosed as unreliable ($\bar{\rho} \approx 61$) and the natural branch is itself noise-dominated. The case marks the applicability boundary discussed in Section 3.5.

Four diagnostics probe the stability and the distributional assumptions of the pipeline on the torsion case. First, a 500-resample bootstrap of the model-form fit of Equation (6) gives a 95% interval of $(0.49, 0.74)$ for the slope $\hat{b}$ and $(9700, 14{,}600)$ for $\bar{\rho}$; the held-out Gaussian-reference coverage is 100% under every bootstrap replicate, so the diagnosis and the band are both stable across residual draws. Replacing the φ^2 regressor with $|\varphi|$ or with the unconstrained quadratic $\varphi + \varphi^2$ changes the extrapolated model-form standard deviation at 240 months from 6.8 N·m to 7.9 N·m and 6.4 N·m, respectively; the chosen form

lies between the alternatives and is the only one that shares the structural time dependence of the propagated covariance.

Second, a Shapiro–Wilk test on the 24 standardized held-out fused residuals gives $W = 0.923$ with $p = 0.069$: no evidence against normality at the 5% level, although the sample is small. The 96 training residuals, in contrast, reject normality decisively ($p < 0.001$), with a lag-one autocorrelation of 0.885, as expected for monthly means sampled from one smooth degradation path. This serial correlation is exactly why the method does not rest interval validity on within-trajectory Gaussian or independence assumptions: validity is established at the population level by the conformal construction, and the AICc selection is confirmed out-of-sample by the held-out bias collapse of Section 3.2, rather than by in-sample likelihood alone.

Third, a sampling-based propagation cross-checks the first-order delta method. Drawing 2×10^5 realizations of the WLS Arrhenius coefficients and pushing them through the exponential map shows that the delta method underestimates the standard deviation of $\hat{B}_{\mathrm{op}}$ by a factor of 1.60 and misses a pronounced right skew (sample skewness 3.5); the trajectory-level sampled interval is correspondingly asymmetric. The fused band, however, is insensitive: substituting the sampled variance for the delta variance changes the fused half-width by less than 10^{-3} percent at every grid point, because σ^2_{MF} exceeds the accelerated sampling variance by three to four orders of magnitude on this case. The delta method is therefore retained in the closed-form pipeline, with the conformal step absorbing any residual distributional error at the population level.

Fourth, a sampled first-passage distribution cross-checks the geometric interval of Equations (14) and (15). Pushing 2×10^4 joint draws of the natural-branch and Arrhenius coefficients through the fused trajectory yields a first-passage distribution with median 137.9 months, skewness 0.24, and 95% interval $(132.8, 143.9)$ months, strictly inside the geometric interval $(125.4, 151.4)$ months. The geometric construction reads the band edges as if they could be attained simultaneously and is therefore conservative, in the operationally safe direction; the sampled interval is reported as the refined alternative, and a full stochastic-process treatment of the crossing density is discussed in Section 4.

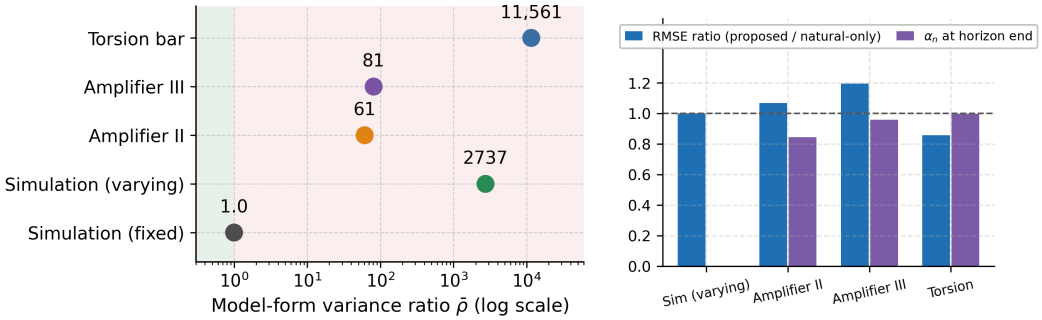
The Rudder II case is a limit of the present formulation: when both branches are individually weak, fusion cannot recover information that neither branch carries. Section 4 returns to these limits in the context of avenues for further work.

4. Discussion

The model-form variance ratio of Equation (8) is a diagnostic for the fusion regime, not merely a variance term. Across the five experimental settings it spans four orders of magnitude: approximately 11,600 on the torsion bar, 81 and 61 on the two amplifier variants, 2700 on the varying-truth simulation whose ground truths contain a hidden secondary stage, and 1 on the fixed-truth control where Arrhenius is exact by construction. Figure 10 visualizes this span on a regime-diagnostic axis. The proposed method's contribution shifts with the ratio. At a large ratio, it is sharp, population-calibrated uncertainty quantification on a fallback-to-natural-only point estimate; at a small ratio, it is precision-weighted point-estimate fusion in the classical sense. A reliability engineer can read the ratio from the training data alone and predict whether fusion will improve point accuracy or only calibration, before any held-out evaluation.

The variance decomposition $\sigma^2_a(t)$ in Equation (5), on its own, under-weights the model-form error that dominates long-distance Arrhenius extrapolation; the OLS-propagated component captures only the sampling uncertainty in the Arrhenius coefficients. Adding the model-form variance of Equation (6) turns the variance term into an empirical reality check on the accelerated branch, by treating natural-storage residuals

against the accelerated prediction as a direct measurement of accelerated-branch error at the operating temperature. The torsion case now identifies the physical origin of that error. The AICc selection finds a secondary relaxation stage worth 0.48 N·m of torque loss per year, consistent with grain-boundary sliding and secondary creep; over the 144 h accelerated window, this stage contributes below the accelerated noise floor and is structurally invisible to the per-temperature regression, yet at decade horizons it dominates the trajectory. The accelerated branch is therefore not poorly fitted but probing the wrong regime: its window can only see primary kinetics, and $\bar{\rho} \gg 1$ is the quantitative signature of the hidden stage. The varying-truth simulation confirms this reading by construction, with a median ratio of 2700 when a secondary stage is present against 0.97 when it is absent. This is the mechanism that produces the safe-fallback behavior visible in Figures 3 and 4, where the proposed method recovers the natural-only trajectory while the four fusion baselines are pulled by the biased accelerated prediction. For test-program design, the same diagnostic could flag accelerated-branch unreliability before fusion, motivating longer accelerated windows or intermediate test temperatures that make the secondary stage identifiable. We caution that the model-form variance is informative about which branch the fusion should trust, not an independent measure of true accelerated-branch accuracy. The natural-storage residuals are themselves the reference, so the term cannot detect misspecification that is symmetric across both branches.



(**a**) Regime diagnostic. (**b**) Cross-case synthesis.

Figure 10. Regime diagram synthesizing the case studies on the model-form variance-ratio axis. Panel (**a**): The five experimental settings span $\bar{\rho}$ over four orders of magnitude, from ≈1 (fixed-truth simulation, Arrhenius exact by construction) through ≈61 and ≈81 (the two amplifier variants) and ≈2700 (varying-truth simulation, hidden secondary stage), to ≈11,600 (torsion bar, severe Arrhenius misspecification at room temperature). Panel (**b**): On each case, the proposed method keeps the held-out RMSE ratio to natural-only at or below about 1.2 (blue bars; below 1 on the torsion case), while the horizon-end natural-branch weight α_n (purple bars) follows the regime that $\bar{\rho}$ predicts, rising toward unity as misspecification grows. The figure synthesizes the headline Discussion insight: $\bar{\rho}$, computed from training data alone, is the single number that predicts which regime fusion will deliver on a new case.

Under the split-conformal construction, validity is guaranteed for every method, and Table 2 shows what each pays for it: the proposed method attains the nominal level with a conformal quantile of $q = 1.01$ and a 0.071 N·m band, while the scalar-weight baselines need quantiles up to 38.7 and bands up to 14.0 N·m. A quantile near unity means the fused $\hat{\sigma}(t)$ was already population-calibrated before the conformal step touched it. The mechanism is the structure of Equations (7) and (11): the additive model-form term makes the fused variance largest exactly where the accelerated branch is least trustworthy, which is the structure a calibrated band requires, and which scalar weights cannot reproduce because they are constants independent of where in time the branches disagree. For reliability decision-making, the conformal guarantee lets engineers use the

lower endpoint T_f^{lo} as a maintenance-scheduling threshold with a stated finite-sample error rate, rather than treating the nominal level as aspirational. The guarantee is relative to the calibration population: coverage transfers to a new case to the extent that the case is exchangeable with the population, which is the assumption that Section 3.5 makes explicit.

The cross-case pattern, summarized in Figure 10, is consistent: the proposed method matches or slightly improves on the natural-only baseline in terms of point accuracy while being the only fusion strategy whose interval is simultaneously valid and sharp. Existing fusion methods cannot achieve either simultaneously because their scalar weights conflate within-branch noise, sampling error, and model-form disagreement into a single residual variance. The proposed method's structural advance is the separation of these three sources, which is what enables the mechanism-aware mean, the time-varying weight, and the calibrated band at once. The implication for multi-fidelity storage fusion is that the relevant design question concerns what the weight should depend on at each time and what variance that dependence implies, rather than what scalar weight summarizes the relative trust between branches.

The proposed method aligns with the relative-entropy weighting of [13] and the Bayesian-calibration-factor framework of [14] on the goal of combining natural and accelerated data into a single posterior on degradation. The agreement ends at the mechanism of weighting. Relative-entropy methods compute a single scalar weight from the divergence between source predictions on the training grid and apply it globally. This is the most direct point of disagreement. Such weights cannot reflect the time-varying mismatch between the accelerated branch's nominal precision and its actual at-room-temperature accuracy. The Bayesian-calibration-factor approach is closer in spirit to the present work, in that it calibrates the accelerated prediction against natural-storage data. The difference is that the BCF calibration enters as a posterior on a mean shift, rather than as an inflation of the predictive variance. This distinction matters for uncertainty quantification specifically. A mean shift cannot widen the predictive band at the points where Arrhenius is least trustworthy, so the resulting nominal variance is far from calibrated, and the conformal step must stretch it by a factor of 6.2 on the varying-truth population, against 1.01 for the proposed method. None of the prior fusion approaches—scalar-weight or Bayesian—includes a finite-sample validity layer; the split-conformal construction adopted here supplies that layer to any of them, which is precisely what makes band sharpness, rather than nominal coverage claims, the fair comparison. Multi-fidelity Gaussian-process methods in adjacent fields, such as the surrogate-based aerodynamic optimization work of [22] and the multi-fidelity Hamilton–Kriging model of [23], use precision-weighted fusion ideas that are mathematically related to the present work. In those settings, the low-fidelity source is a coarser simulator rather than an extrapolated Arrhenius model, and the source of model-form error is qualitatively different. The additive model-form variance in the present work is the structural addition that adapts the multi-fidelity precision-weighting idea to the storage-data setting, where extrapolation distance dominates; the present framework is regression-based throughout and claims no kernel Gaussian-process machinery.

Six specific limitations bound the present work: First, the method assumes a single dominant mechanism per case, with the nested family covering only that mechanism's primary and secondary stages; the per-temperature regression cannot represent failures in which the dominant degradation mode itself changes at high test temperatures, and the model-form variance would catch the symptom as an inflated residual without offering a mechanism-discovery procedure. Second, single-trajectory extrapolation has a finite honest horizon: the window-sensitivity study shows the Gaussian-reference coverage falling to 12.5% when the held-out window reaches two thirds of the training window, because the secondary-stage slope identified from the shortened window under-predicts the later

curvature. The band is honest rather than inflated, and the decay is the visible price; validity claims are therefore made at the population level. Third, the conformal guarantee is relative to its calibration population: transfer to a new field case assumes exchangeability with that population, and the simulated population, however varied, is a model of the field rather than the field itself. When fleet data from multiple stored components become available, the same construction calibrates directly on real profiles. Fourth, the model-form variance uses natural-training residuals as the reference for accelerated-branch error and, therefore, cannot detect misspecifications that are symmetric across both branches; independent residuals from leave-one-temperature-out cross-validation would be a stronger diagnostic and are a planned extension. Fifth, the first-passage interval is a band-crossing construction; the sampled-trajectory cross-check shows it to be conservative relative to the parameter-uncertainty first-passage distribution, and a full stochastic-process treatment, modeling the crossing density of a Wiener or gamma degradation process with path-dependent correlation, remains an avenue for future work. Sixth, the delta-method propagation underestimates the sampled standard deviation of $\hat{B}_{\mathrm{op}}$ by a factor of 1.6 and misses its right skew; on the present cases, this is immaterial because the model-form variance dominates by orders of magnitude, but in low-$\bar{\rho}$ regimes the sampling-based propagation quantified in Section 3.5 should replace the delta step.

The variance-decomposition and model-form-variance concepts generalize beyond Arrhenius storage to any multi-fidelity setting where a high-fidelity but data-poor source and a low-fidelity but data-rich source are combined and the low-fidelity source carries time-varying model-form error. For practitioners, the ratio $\bar{\rho}$ identified in D1 is a one-number diagnostic that can accompany any storage-fusion analysis without changing the existing reliability-engineering workflow, and the conformal layer can be added to any existing fusion pipeline to convert nominal intervals into guaranteed ones. Several extensions follow from the present results. A multi-mechanism extension of the per-temperature regression would relax the single-mechanism assumption; a stochastic-process first-passage model would replace the band-crossing construction; and conformal calibration on real fleet populations would replace the simulated calibration set as such data accumulate. Section 5 returns to these as the concluding statement.

5. Conclusions

The reliability of aerospace dormant components is increasingly inferred from a fusion of natural-storage and accelerated-storage data. Existing fusion methods, however, are pulled away from natural-storage observations whenever the underlying Arrhenius extrapolation is misspecified at the operating temperature. This paper presents an empirically calibrated multi-fidelity fusion that selects a mechanism-specific natural-branch model by AICc; decomposes the accelerated-branch variance into intrinsic, sampling, and model-form components, with the last fitted additively from natural-storage residuals; and calibrates the fused band by split-conformal prediction. Three findings were established: First, the mechanism-specific selection and the additive model-form variance together produced safe-fallback behavior on the two real case studies. On the torsion bar, AICc decisively identified a secondary relaxation stage invisible to the accelerated window, and the held-out RMSE was 0.045 N·m for the proposed method against 0.052 N·m for natural-only, while the four alternative fusion methods produced values from 2.8 to 126 times larger across the two cases; the fused mean crossed the failure threshold at 11.5 years, with a band-crossing interval of 10.4 to 12.6 years. Second, on a 1000-run varying-truth Monte Carlo population with disjoint calibration and test halves, the conformally calibrated band attained 95.5% out-of-sample trajectory coverage, as guaranteed, at a mean width of 0.071 N·m, 97 to 196 times narrower than the other fusion strategies

at the same level, with a conformal quantile of 1.01, indicating that the fused variance was already population-calibrated. Third, the model-form variance ratio acted as a single-number diagnostic for the fusion regime, spanning four orders of magnitude from ≈ 1 where Arrhenius is exact to $\approx$ 11,600 on the torsion bar. The ratio identifies in advance whether the proposed method's contribution is calibrated uncertainty on a fallback estimate, or precision-weighted point-estimate fusion. The transferable insight is that a mechanism-aware mean family plus an additive model-form variance lets a multi-fidelity fusion detect its own misspecification, turning the fusion weight into a falsifiable engineering diagnostic with a validity layer that any fusion pipeline can adopt. The framework is limited to single-mechanism regimes, population-relative conformal guarantees, and a band-crossing first-passage construction. Three immediate extensions are natural next steps: a multi-mechanism per-temperature regression, a stochastic-process first-passage model with path-dependent correlation, and conformal calibration on real fleet populations as such data accumulate.

Author Contributions: Conceptualization, S.Z., C.-W.F., and B.Z.; methodology, S.Z. and B.Z.; software, S.Z.; investigation, S.Z. and S.R.; formal analysis, S.Z. and S.R.; data curation, S.R., H.L., and S.H.; resources, H.L., S.H., and C.-W.F.; validation, S.R., H.L., and S.H.; writing—original draft preparation, S.Z.; writing—review and editing, S.R., H.L., S.H., C.-W.F., and B.Z.; visualization, S.Z.; supervision, C.-W.F. and B.Z.; project administration, C.-W.F.; funding acquisition, C.-W.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-supported by the National Natural Science Foundation of China (Grant No. 52375237) and the National Science and Technology Major Project (Grant No. J2022-IV-0012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The natural-storage and accelerated-storage measurement data for the two real case studies are subject to internal aerospace test-program confidentiality and are available from the corresponding author upon reasonable request, conditional on the program's data-release policy. Aggregated and de-identified summary statistics, the synthetic Monte Carlo simulation, and the Python code used for figure generation are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank the funding agencies listed above for their support, as well as the colleagues at the Aerospace Science and Industry Defense Technology Research Testing Center for facilitating the field-monitoring records used in the case studies. During the preparation of this manuscript, the authors used Claude (Anthropic) for language editing and figure-code drafting assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations and mathematical symbols are used in this manuscript:

Acronyms.

ADT	Accelerated degradation testing
BCF	Bayesian calibration factor
CI	Credible interval
FPT	First-passage time
GP	Gaussian process
KL	Kullback–Leibler (divergence)
MC	Monte Carlo
AICc	Corrected Akaike information criterion

OLS	Ordinary least squares
RMSE	Root-mean-square error
RUL	Remaining useful life
SCP	Split-conformal prediction
WLS	Weighted least squares
Mathematical symbols.	
α	Significance level of the credible interval (0.05 for a 95% CI)
$\alpha_n(t)$	Time-varying natural-branch fusion weight
$A(T), B(T)$	Per-temperature degradation-model coefficients
E_a	Arrhenius activation energy (eV)
K	Number of accelerated-storage test temperatures
$P(t)$	Degradation parameter at storage time t
P_0	Initial value of the degradation parameter
P_{fail}	Failure threshold on $P(t)$
R	Universal gas constant
$r(t)$	Natural-storage residual against the accelerated prediction
T_{op}	Operating temperature
T_k	k-th accelerated-storage test temperature, $T_k > T_{\text{op}}$
T_f	First-passage time, $\inf\{t : P(t) \leq P_{\text{fail}}\}$
$T_f^{\text{lo}}, T_f^{\text{hi}}$	Lower/upper endpoints of the credible interval on T_f
$\phi(t)$	Family-specific time profile ($\ln t$ stress relaxation; t linear drift)
$\sigma^2_{\text{intrinsic}}$	Pooled intrinsic noise variance of the accelerated branch
$\sigma^2_{\text{extrap,OLS}}(t)$	OLS-propagated extrapolation-variance component
$\sigma^2_{\text{MF}}(t)$	Additive empirical model-form variance
$\bar{\rho}$	Mean model-form-to-OLS variance ratio (regime diagnostic)
q	Split-conformal calibration quantile
$\sigma^2_a(t)$	Accelerated-branch predictive variance at T_{op}
$\hat{P}(t)$	Fused trajectory at T_{op}
$\hat{\sigma}^2(t)$	Fused predictive variance at T_{op}
$\mathbf{Y}^{(n)}$	Natural-storage data vector
$\mathbf{Y}^{(a,k)}$	Accelerated-storage data vector at temperature T_k

References

1. Zio, E. Prognostics and health management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliab. Eng. Syst. Saf.* **2022**, *218*, 108119. [CrossRef]
2. 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]
3. Zhao, H.; Wang, L.; Zhang, L.; Bian, B. Numerical prediction of fatigue life for landing gear considering the shock absorber travel. *Aerospace* **2025**, *12*, 42. [CrossRef]
4. Wang, L.; Zhao, H.; Zhang, L.; Bian, B. Fatigue life prediction and experimental study of landing gear components via FKM local stress approach. *Aerospace* **2025**, *12*, 1026. [CrossRef]
5. Sun, Y.; Feng, H.; Zheng, B.; Wen, J.; Chao, A.; Fei, C.W. Multi-agent reinforcement symbolic regression for the fatigue life prediction of aircraft landing gear. *Aerospace* **2025**, *12*, 718. [CrossRef]
6. Zhang, S.; Cao, Y.; Li, T.; Ma, X.; Li, H. Osprey optimization algorithm-optimized Kriging-RBF method for radial deformation reliability analysis of compressor blade angle crack. *Aerospace* **2025**, *12*, 867. [CrossRef]
7. Cubillo, A.; Vermeulen, R.; De Rovere, M.; Megantoro, P.; Marín, S.V. An integrated monitoring, diagnostics, and prognostics system for aero-engines under long-term performance deterioration. *Aerospace* **2024**, *11*, 217. [CrossRef]
8. Abdulai, S.; Hong, S.; Hayes, C.; Marquez, S.; Kaul, P. A framework for an ML-based predictive turbofan engine health model. *Aerospace* **2025**, *12*, 725. [CrossRef]
9. Sun, Y.; Wen, J.; Li, J.; Chao, A.; Fei, C.W. Novel integrated model approach for high cycle fatigue life and reliability assessment of helicopter flange structures. *Aerospace* **2025**, *12*, 78. [CrossRef]
10. Lin, P.; Zhao, T.; Xie, Y.; Yuan, M.; Li, X. Acoustic and vibration response and fatigue life analysis of thin-walled connection structures under heat flow conditions. *Aerospace* **2024**, *11*, 287. [CrossRef]

11. Sun, Y.; Cai, B.; Liu, J. Storage reliability assessment method for aerospace electromagnetic relay based on belief reliability theory. *Appl. Sci.* **2022**, *12*, 8637. [CrossRef]
12. Meeker, W.Q.; Escobar, L.A. *Statistical Methods for Reliability Data*; John Wiley & Sons: New York, NY, USA, 1998.
13. Zhang, S.; Ni, R.; Xu, R.; Ma, X.; Li, H. A Multi-Stage Storage Data Fusion Evaluation Method Based on Relative Entropy Weight Combination. *Equip. Environ. Eng.* **2024**, *21*, 1–8.
14. Wang, L.; Pan, R.; Li, X.; Jiang, T. A Bayesian reliability evaluation method with integrated accelerated degradation testing and field information. *Reliab. Eng. Syst. Saf.* **2013**, *112*, 38–47. [CrossRef]
15. Lv, Y.; Chen, X.; Li, Y.; Tian, Y.; Zhang, F. Reliability evaluation of landing gear retraction/extension accuracy based on Bayesian theory. *Aerospace* **2025**, *12*, 300. [CrossRef]
16. Pang, Z.; Si, X.; Hu, C.; Du, D.; Pei, H. A Bayesian inference for remaining useful life estimation by fusing accelerated degradation data and condition monitoring data. *Reliab. Eng. Syst. Saf.* **2021**, *208*, 107341. [CrossRef]
17. Ma, Z.; Wang, S.; Ruiz, C.; Zhang, C.; Liao, H.; Pohl, E. Reliability estimation from two types of accelerated testing data considering measurement error. *Reliab. Eng. Syst. Saf.* **2020**, *193*, 106610. [CrossRef]
18. Liu, D.; Wang, S.; Zhang, C.; Tomovic, M. Bayesian model averaging based reliability analysis method for monotonic degradation dataset based on inverse Gaussian process and Gamma process. *Reliab. Eng. Syst. Saf.* **2018**, *180*, 25–38. [CrossRef]
19. Kennedy, M.C.; O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **2000**, *87*, 1–13. [CrossRef]
20. Brevault, L.; Balesdent, M.; Hebbal, A. Overview of Gaussian-process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerosp. Sci. Technol.* **2020**, *107*, 106339. [CrossRef]
21. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
22. Garbo, A.; Parekh, J.; Rischmann, T.; Bekemeyer, P. Multi-fidelity adaptive sampling for surrogate-based optimization and uncertainty quantification. *Aerospace* **2024**, *11*, 448. [CrossRef]
23. Zhang, S.; Ma, J. Adaptive sequential infill sampling method for experimental optimization with multi-fidelity Hamilton Kriging model. *Aerospace* **2025**, *12*, 913. [CrossRef]
24. Wen, J.; Sun, Y.; Chao, A.; Zheng, B.; Li, J.; Feng, H. Deep-reinforcement-learning-enhanced Kriging modeling method with limit state dominant sampling for aeroengine structural reliability analysis. *Aerospace* **2025**, *12*, 752. [CrossRef]
25. Liu, X.; Deng, J.; Chen, H.; Zhai, G.; Wu, J. An efficient and multi-fidelity reliability-based design optimization method based on a novel surrogate model local update strategy. *Comput. Methods Appl. Mech. Eng.* **2024**, *430*, 117219. [CrossRef]
26. Kapusuzoglu, B.; Mahadevan, S. Adaptive reliability analysis for multi-fidelity models using a collective learning strategy. *Struct. Multidiscip. Optim.* **2021**, *64*, 3993–4016. [CrossRef]
27. Ravi, K.; Fediukov, V.; Dietrich, F.; Neckel, T.; Buse, F.; Bergmann, M.; Bungartz, H.J. Multi-fidelity Gaussian process surrogate modeling for regression problems in physics. *Mach. Learn. Sci. Technol.* **2024**, *5*, 045015. [CrossRef]
28. Zhang, Z.; Si, X.; Hu, C.; Lei, Y. Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods. *Eur. J. Oper. Res.* **2018**, *271*, 775–796. [CrossRef]
29. Ye, Z.S.; Chen, N.; Shen, Y. A new class of Wiener process models for degradation analysis. *Reliab. Eng. Syst. Saf.* **2015**, *139*, 58–67. [CrossRef]
30. Yan, B.; Ma, X.; Yang, L.; Wang, H.; Wu, T. A novel degradation-rate-volatility related effect Wiener process model with its extension to accelerated ageing data analysis. *Reliab. Eng. Syst. Saf.* **2020**, *204*, 107138. [CrossRef]
31. Sankararaman, S.; Mahadevan, S. Likelihood-based representation of epistemic uncertainty due to sparse point data and/or interval data. *Reliab. Eng. Syst. Saf.* **2011**, *96*, 814–824. [CrossRef]
32. Nabarro, F.R.N.; de Villiers, H.L. *The Physics of Creep: Creep and Creep-Resistant Alloys*; Taylor & Francis: London, UK, 1995.
33. Grasser, T. Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities. *Microelectron. Reliab.* **2012**, *52*, 39–70. [CrossRef]
34. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [CrossRef]
35. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed.; Springer: New York, NY, USA, 2002.
36. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005.
37. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
38. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [CrossRef]