Uncertainty Discourse and Applications

www.uda.reapress.com

Uncert. Disc. Appl. Vol. 3, No. 4 (2026) 296–306.

Paper Type: Original Article

Coverage of Deep Remaining Useful Life Prediction Intervals under C MAPSS Cross Subset Transfer

David R Harrison[1,*], Aisha B Saleh[2], Vikram Iyengar[3], Lea Hoffmann[4]

[1] Department of Computer Science, University of Cape Town, Cape Town, South Africa; d.r.harrison@cs.uct.ac.za.

[2] College of Engineering, Qatar University, Doha, Qatar.

[3] Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India.

[4] CISPA Helmholtz Center for Information Security, Saarbrucken, Germany.

Citation:

Received: 20 January 2026 *Revised: 15 May 2026* *Accepted: 12 August 2026*	Harrison, D. R., Saleh, A. B., Iyengar, V., & Hoffmann., L. (2026). Coverage of deep remaining useful life prediction intervals under C MAPSS cross subset transfer. *Uncertainty Discourse and Applications*, *3*(4), 296-306.

Abstract

Conformal prediction intervals require a calibration–deployment relationship that can fail when a Remaining Useful Life (RUL) model transfers between operating conditions or fault modes. This study compares seven interval strategies on three compact backbones across all twelve ordered transfers among the four public Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) subsets, giving thirty-six transfer blocks. The reported pooled split-conformal coverage decreases from 0.918 in-domain to 0.606 under transfer, with a minimum of 0.083. Density-Ratio-Weighted Calibration (PORTCAL) has mean coverage 0.589 and Winkler score 358.2 cycles, compared with 0.606 and 347.7 for pooled calibration. A feedback-based, clipped Adaptive Conformal Inference (ACI) implementation has the lowest reported mean Winkler score, 221.1, with coverage 0.755 and width 74.8 cycles. MC-dropout and ensembles have higher transfer coverage than offline conformal variants but remain below the nominal 0.90. These observations concern a simulated benchmark and the specific implementations. Clipping in ACI, estimated and capped weights, dependent trajectory windows and immediate retrospective RUL feedback limit theoretical and deployment interpretations. The results support separate reporting of in-domain and transfer coverage, accompanied by interval width and an explicit information contract.

Keywords: Remaining useful life, Conformal prediction, Distribution shift, Uncertainty quantification, Predictive maintenance, Commercial modular aero-propulsion system simulation, Coverage audit.

1 | Introduction

Remaining Useful Life (RUL) prediction estimates the operating cycles remaining before failure from a degrading asset's telemetry. Convolutional predictors [1] and transformer designs [2] have been studied on Commercial Modular Aero-Propulsion System Simulation (C-MAPSS), a simulated turbofan benchmark [3],

Corresponding Author: d.r.harrison@cs.uct.ac.za

https://doi.org/10.48313/uda.v3i4.116

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

[4]. Uncertainty methods include dropout-based approximate inference [5], deep ensembles [6] and conformal intervals [7], [8]. Coverage and width are relevant to maintenance planning, but empirical interval performance is distinct from demonstrated economic or safety benefit.

Split-conformal prediction gives marginal coverage under exchangeability of calibration and future scores, with a model trained separately [8]. C-MAPSS contains subsets with different operating-condition and fault-mode structures [4], providing a controlled transfer setting. Importance-weighted conformal methods address covariate shift under an invariant conditional response law and suitable likelihood-ratio weights [9]. Online Adaptive Conformal Inference (ACI) instead updates its level from realized error feedback [10]; later work also adapts the step size [11]. These theories provide distinct information requirements and do not automatically validate their modified benchmark implementations.

The study asks three questions. How does empirical interval coverage change between in-domain and cross-subset tasks? Does the tested Density-Ratio-Weighted Conformal Calibration (PORTCAL) implementation improve transfer coverage and Winkler score over unweighted and condition-based calibration? How much does the tested feedback-based ACI replay improve those metrics, and at what width cost? The third question assumes RUL feedback is supplied after each interval in a retrospective replay; it is not a label-free deployment setting.

The comparison combines twelve ordered transfer pairs with three backbones, while retaining four in-domain tasks for each backbone. The reported offline conformal coverage falls substantially under transfer, PORTCAL does not improve the aggregate endpoints, and ACI replay improves mean coverage while remaining below nominal. The contribution is this implementation-level comparison under a compact training protocol. It neither diagnoses the shift uniquely nor establishes that one interval strategy is generally valid across fleets.

Sections 2–4 establish related methods, notation and implementation choices. Section 5 defines the benchmark protocol. Section 6 reports coverage, interval scores, component comparisons and point-prediction context, followed by interpretation and limitations.

2 | Related Work

2.1 | Deep Prognostics on C-MAPSS

Deep RUL pipelines map windows of sensor observations to a target clipped at an early-life plateau. Convolutional models [1] and dual-aspect transformer architectures [2] provide established designs. The backbones in this study are deliberately compact: their role is to hold the predictor fixed for comparisons among conformal wrappers, rather than reproduce the exact published architectures or establish a new point-prediction state of the art.

2.2 | Uncertainty Quantification for RUL

Dropout-based approximate inference [5] and ensembles [6] estimate uncertainty through stochastic or multi-model predictions. Conformal RUL methods convert a point predictor into an interval using calibration residuals [7]. These constructions should be distinguished from the present transfer experiment: a source's stated coverage guarantee applies under its assumptions, and a transfer failure cannot by itself show that prior work claimed validity under the same shift.

2.3 | Conformal Validity under Shift

Weighted conformal prediction retains coverage under an appropriate weighted exchangeability model, including covariate shift with known likelihood ratio and unchanged conditional response law [9]. Estimated ratios introduce additional uncertainty. ACI uses error feedback to control empirical error frequency under an unprojected update and defined extreme quantiles [10]; the later method in [11] changes the adaptation

rule. The tested implementations include clipping and finite safeguards, so their coverage is evaluated empirically.

3 | Problem Setting and Preliminaries

Each engine unit i supplies a sensor trajectory. Units are assigned separately to source training, source calibration and official test roles. At cycle t, a window $X_{it} \in \mathbb{R}^{(w \times m)}$ contains w = 30 observations in m = 14 selected sensor channels, with stride two. For a complete training trajectory of length T_i and zero-based index t, the study uses

$$y_{it} = \min\{130, \max\{0, T_i - 1 - t\}\}, \quad t = 0,\ldots,T_i-1. \tag{1}$$

Each backbone f_theta outputs a Gaussian predictive head (mu, log sigma^2) trained with the negative log-likelihood

$$L(\text{theta}) = (1/n)\ \text{SUM_j}\ [\ 0.5 \log \text{sigma}^2(X_j) + 0.5\ (y_j - \text{mu}(X_j))^2 / \text{sigma}^2(X_j)\ ]. \tag{2}$$

The fitted Gaussian head provides a mean μ(X) and positive scale σ(X). Normalized conformal intervals have the form [μ(X)−qσ(X), μ(X)+qσ(X)], whereas absolute-residual conformal uses [μ(X)−q, μ(X)+q]. Coverage is the fraction of evaluated labels contained in the interval, and mean width summarizes interval size. The Winkler score combines width and missed-label penalties at α = 0.10:

$$W(y, l, u) = (u - l) + (2/\text{alpha})(l - y)1\{y < l\} + (2/\text{alpha})(y - u)1\{y > u\}. \tag{3}$$

The ordinary split-conformal theorem concerns exchangeable calibration and test scores, conditional on a separately trained model [8]. Pooling overlapping windows from the same trajectories does not establish that assumption. Cross-subset tasks additionally change the calibration–deployment relationship. The reported coverage drift is the in-domain mean minus the transfer mean:

$$\text{Delta} = \text{cov_in-domain(SC)} - \text{cov_transfer(SC)}. \tag{4}$$

This drift summarizes the change in empirical coverage. In-domain tasks calibrate and deploy within one subset, using held-out official test trajectories; transfer tasks calibrate on source-subset units and deploy on the target subset's official test trajectories.

4 | Calibration and Adaptation Strategies under Audit

Seven interval strategies are evaluated for each backbone and source subset. The conformal wrappers and ACI reuse the seed-101 predictor and source calibration data. MC-dropout averages stochastic outputs, and the ensemble averages five trained members; their prediction centers can therefore differ from seed-101. They share a training design rather than an identical point predictor.

4.1 | Split Conformal and Variants

Pooled Split Conformal (SC) computes normalized nonconformity scores on the calibration split,

$$s_j = |y_j - \text{mu}(X_j)| / \text{sigma}(X_j). \tag{5}$$

For SC, q is the k-th sorted normalized score, $k = \lceil (n+1)(1-\alpha) \rceil$; if k > n, the theoretical finite-sample convention uses q = ∞. SC-ABS uses absolute residuals and a half-width q without multiplying again by σ(X). Mondrian calibration assigns windows to K = 6 operational-setting clusters fitted on source training rows and uses a cluster-specific quantile. Empty or small clusters require an explicit fallback. For single-condition subsets the clustering may be uninformative, but an exact reduction to SC depends on the implemented labels and fallback. No conditional coverage guarantee is established for these dependent, shifted windows.

4.2 | Density-Ratio-Weighted Conformal Calibration

PORTCAL estimates a target-to-source weight from a logistic classifier fitted to standardized per-window channel means. Let c(x) denote the fitted target-class probability and πT, πS the class proportions used to fit that classifier. Bayes' rule gives the density-ratio relationship

$$pT(x)/pS(x) = (\pi S/\pi T)\, c(x)/(1-c(x)); \text{ the reported implementation uses odds } \hat{w}(x) = c(x)/(1-c(x)). \quad (6)$$

Without clipping, a common weight factor cancels from the weighted quantile, so odds suffice [9]. Clipping before this cancellation can make the fitted class proportions relevant. For calibration weights w_i and test-point weight $\hat{w}(x)$, the reported threshold is

$$q_hat(x) = \inf\{ s : SUM_i\ w_i\ 1\{s_i <= s\} >= (1 - alpha)(SUM_i\ w_i + w_hat(x)) \}. \quad (7)$$

The search uses cumulative calibration weights, clipped to [0.01, 100]. The theoretical weighted distribution includes test-point mass at infinity; if the finite scores do not reach the required mass, q should be infinite. The implementation instead caps that threshold at 1.5 times the largest calibration score. This is a practical modification without the oracle guarantee. PORTCAL consumes unlabeled target features, while both its ratio accuracy and the covariate-shift assumption require separate assessment.

4.3 | Adaptive Conformal Inference

ACI [10] processes windows ordered by unit and cycle. Each interval is issued before its label is supplied in the retrospective replay, after which the miss indicator updates the level:

$$alpha_\{t+1\} = alpha_t + gamma\ (alpha - err_t),\ err_t = 1\{y_t \notin C_t\},\ gamma = 0.02, \quad (8)$$

The implementation projects α_t to $[10^{-4}, 0.5]$ and selects a quantile from fixed source calibration scores. This is a clipped ACI variant. The distribution-free frequency bound for the unprojected recursion relies on its update identity and extreme quantile conventions [10]; projection and finite score support prevent invoking that bound here. The simulation of immediate feedback also differs from maintenance deployment, where exact cycle-wise RUL is generally resolved retrospectively after failure.

4.4 | Uncalibrated Baselines

MC-dropout performs twenty stochastic forward passes with dropout active [5]. The ensemble averages five independently seeded members [6]. Their Gaussian intervals use mean prediction $\pm 1.645\sqrt{(E[\sigma^2]+Var[\mu])}$, combining within-member predictive variance with between-prediction variance. They use no calibration set and have no conformal coverage guarantee. Their finite Gaussian intervals test a separate uncertainty design, with different centers and additional inference or training cost.

5 | Experimental Protocol

The experiment uses the public C-MAPSS distribution, which consists of simulated turbofan trajectories with operational settings, noisy sensors and official final-cycle RUL labels [4]. The experiment reports retrieval on October 9, 2026. Windows have length thirty and stride two, with padding for early cycles and fourteen selected sensors. The reported source split allocates about 70% of units to training and 30% to calibration. Exact unit IDs are needed because the stated every-tenth-unit pattern does not uniquely specify that allocation. Z-score statistics are fitted to source training windows and reused for calibration and deployment.

Table 1. Reported C-MAPSS counts after the study's windowing and unit split. These are processed counts, not a substitute for an archive checksum and unit-selection manifest.

Subset	Train Units	Cal Units	Test Units	Train Win	Cal Win	Test Win	Final-RUL Range
FD001	70	30	100	7227	3114	6571	7-145
FD002	182	78	259	18685	8257	16990	6-194
FD003	70	30	100	9022	3365	8320	6-145
FD004	174	75	248	20971	9721	20539	6-195

Three compact backbones are trained for each source subset: a two-stage one-dimensional convolutional model, a bidirectional GRU and a two-layer transformer. The first and third are motivated by [1], [2], but do not reproduce those architectures exactly. Each has the Gaussian head of *Eq. (2)*, Adam learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, eight fixed epochs, gradient clipping at five and target scaling by 130. Five seeds, 101–105, give sixty models. Seed 101 supplies SC, SC-ABS, Mondrian, PORTCAL, ACI and MC-dropout; the ensemble uses all five members. This provides a common backbone design, with the ensemble using five times as many trained members.

Unit-level separation limits direct reuse of a trajectory between training and calibration, and source-only scaling avoids fitting normalization to target labels. These choices do not make overlapping windows statistically independent or ensure exchangeability of window-level calibration scores. The experiment reports sixty checkpoints, forty-eight evaluation records and mean training time of 19.8 seconds on two CPU threads. Verification requires the unit manifest, environment, hardware and run records.

Table 2. The seven audited wrappers (SC-ABS is the ablation of SC) and their information requirements.

Wrapper	Inputs	Target-Side Labels Used	Adaptation
SC / SC-ABS	Calibration residuals	None	Offline
Mondrian	Calibration residuals + op-settings clusters	None	Offline, conditional
PORTCAL	Calibration residuals + unlabeled target features	None	Offline, transductive
ACI	Calibration scores + realized online feedback	Labels as they arrive	Online
MC-dropout	Stochastic forward passes	None	None
Ensemble	Five seed members	None	None

Twelve ordered source-to-target pairs crossed with three backbones define thirty-six transfer blocks; four in-domain tasks per backbone give twelve references. The reported primary endpoint is mean Winkler score over deployment windows. Friedman and Nemenyi analyses summarize rankings, and Wilcoxon comparisons against PORTCAL use Holm correction [12], [13]. These blocks share source models and target trajectories, limiting an independent-block interpretation. The study provides no accessible dated registration, so the analysis is described as specified rather than preregistered. Effect sizes and endpoint-specific ordering accompany the retained reported p-values.

6 | Results

6.1 | RQ1: The Coverage Audit

Table 3 reports pooled SC coverage of 0.918 across twelve in-domain tasks and 0.606 across thirty-six transfer blocks, a difference of 0.312. The lowest reported block is DCNN FD002→FD003 at 0.083, and BiGRU FD004→FD002 reaches 0.915. Thousands of windows do not by themselves imply tight uncertainty because windows overlap within units. Mondrian coverage is 0.605. These observations show a transfer failure for the tested calibration schemes; they do not identify which distribution-shift component caused it.

Averaging across backbones, SC coverage is reported as 0.302 for FD003→FD004, 0.315 for FD003→FD002 and 0.377 for FD002→FD003. Higher values include 0.901 for FD004→FD002, 0.836 for FD002→FD004 and 0.829 for FD001→FD003. NASA describes FD001/FD002 as one fault mode and

FD003/FD004 as two, with one operating condition for FD001/FD003 and six for FD002/FD004 [4]. Some low-coverage pairs cross both axes, while FD003→FD004 shares fault-mode count. The pattern is heterogeneous and does not establish a uniquely dominant concept-shift explanation.

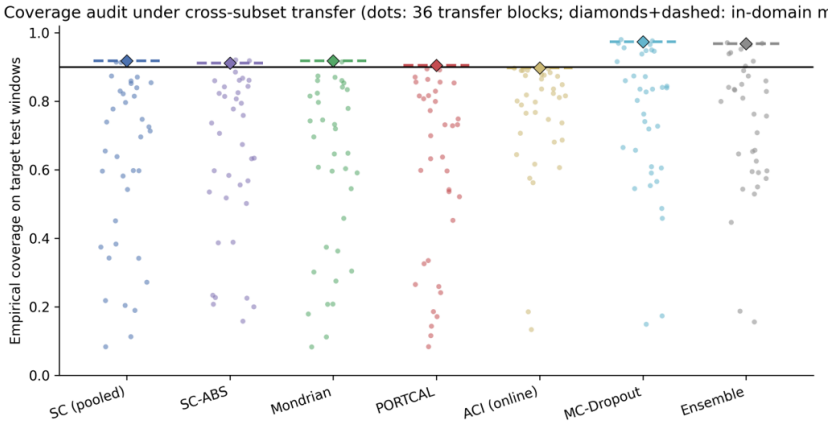


Fig. 1. Reported coverage of nominal 90% intervals.

Note (*Fig. 1*): Dots represent thirty-six transfer blocks; diamonds and dashed lines indicate in-domain means, and the horizontal line is 0.90. The original image's long title is cropped at the right edge.

MC-dropout and ensembles have reported mean transfer coverage of 0.749 and 0.737, with in-domain coverage of 0.974 and 0.968. Their widths are 87.7 and 80.8 cycles, respectively, compared with 52.0 for SC: approximately 68.7% and 55.4% wider. They remain below nominal transfer coverage. Higher coverage and width do not establish that predicted variance reliably increases with unfamiliar inputs, or that the resulting intervals are safe for maintenance decisions.

Table 3. Interval-method aggregates over the 36 transfer blocks (per-block means over deployment windows; Winkler and width in cycles; lower is better except coverage fidelity to 0.90).

Wrapper	Winkler (Mean±sd)	Coverage	|cov-0.9|	Width (Cycles)	In-Domain Cov.	Drift Δ
SC (pooled)	347.7±302.9	0.606±0.253	0.296±0.251	52.0±12.4	0.918	+0.312
SC-ABS	310.6±223.6	0.631±0.239	0.271±0.237	64.3±13.0	0.912	+0.281
Mondrian	348.8±303.3	0.605±0.257	0.297±0.255	52.0±12.9	0.918	+0.314
PORTCAL	358.2±303.9	0.589±0.265	0.311±0.265	49.8±12.8	0.905	+0.316
ACI (online)	221.1±236.5	0.755±0.176	0.145±0.176	74.8±22.4	0.897	+0.142
MC-Dropout	260.6±217.5	0.749±0.211	0.181±0.185	87.7±12.3	0.974	+0.224
Ensemble	278.9±273.3	0.737±0.206	0.187±0.184	80.8±19.8	0.968	+0.232

6.2 | RQ2: Offline Calibration under Transfer

The experiment reports Friedman statistics $\chi^2(6) = 102.0$ for Winkler score and $\chi^2(6) = 91.1$ for absolute coverage error, with p-values 9.6×10^{-20} and 1.8×10^{-17}. *Fig. 3* places ACI at mean rank 2.69, PORTCAL at 3.19, MC-dropout at 3.25, ensembles at 3.56, Mondrian at 4.86, SC at 4.92 and SC-ABS at 5.53. Aggregate Winkler means order the methods differently: ACI is lowest at 221.1 and PORTCAL highest at 358.2. Rank and mean ordering can differ, but the corresponding paired results require reconciliation with the underlying per-block records. The reported critical difference also appears to omit the standard $\sqrt{2}$ normalization, so no Nemenyi separation is certified.

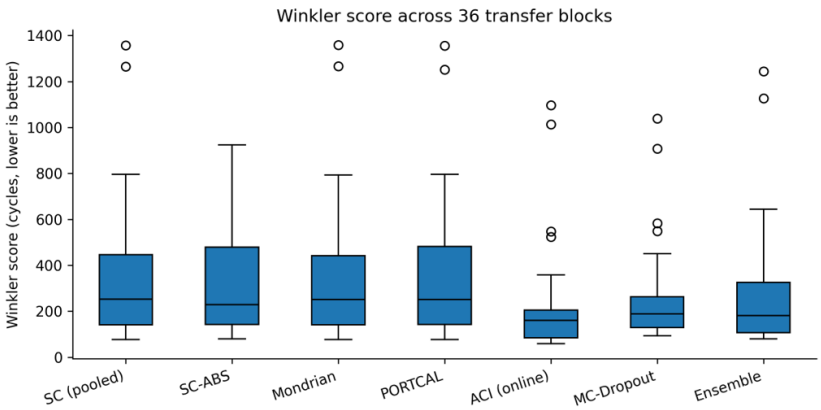


Fig. 2. Winkler-score distributions across the 36 transfer blocks per wrapper (lower is better).

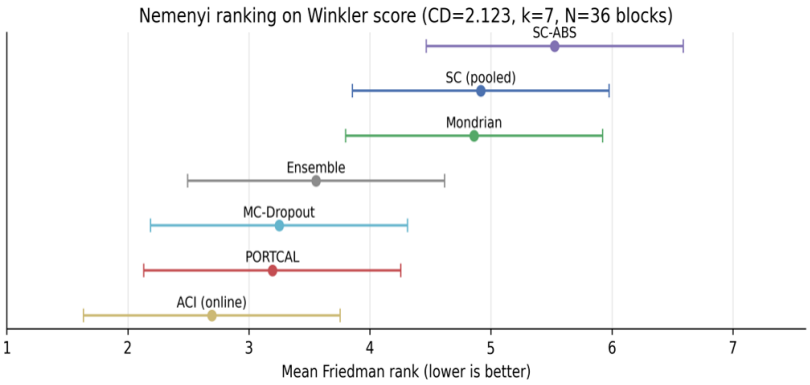


Fig. 3. Reported mean-rank plot for seven wrappers over thirty-six blocks.

Note (*Fig. 3*): The embedded CD is 2.123; the conventional Nemenyi formula uses a studentized-range quantile divided by √2. Rankings and paired mean-score summaries require reconciliation before inferential interpretation.

Table 4 reports PORTCAL-minus-alternative Winkler differences of +137.1 cycles against ACI, +97.5 against MC-dropout and +79.3 against ensembles, with reported Holm-adjusted p-values 1.75×10^{-10}, 4.54×10^{-6} and 4.38×10^{-5}. Against SC, SC-ABS and Mondrian the differences are +10.5, +47.5 and +9.4 cycles, respectively. All retained adjusted p-values in that table are below 0.05. Their calibration depends on the block structure and paired test implementation. *Fig. 4* depicts the width–coverage-error trade-off rather than a guarantee of coverage.

The experiment reports mean target weights of 9.60±6.30 across transfer blocks, a range of 2.21–23.24 and average within-task 90th/10th-percentile ratio of 117. These diagnostics indicate nonconstant fitted weights, not verified likelihood-ratio accuracy or sufficient support overlap. The experiment reports a zero capped-threshold fraction in every block. Weighted intervals are narrower on average than SC, 49.8 versus 52.0 cycles, and their coverage is lower. Poor ratio estimation, limited feature summaries, clipping, score dependence and conditional shift are competing explanations; the experiment does not isolate them.

Table 4. Paired Wilcoxon signed-rank tests over the 36 blocks (positive Δ = PORTCAL worse).

Comparison (PORTCAL - Aalternative)	Mean ΔWinkler (Cycles)	Wilcoxon p (Holm-Corrected)
PORTCAL vs SC (pooled)	+10.5	1.63e-03
PORTCAL vs SC-ABS	+47.5	1.10e-02
PORTCAL vs mondrian	+9.4	4.38e-05
PORTCAL vs ACI (online)	+137.1	1.75e-10
PORTCAL vs MC-dropout	+97.5	4.54e-06
PORTCAL vs ensemble	+79.3	4.38e-05

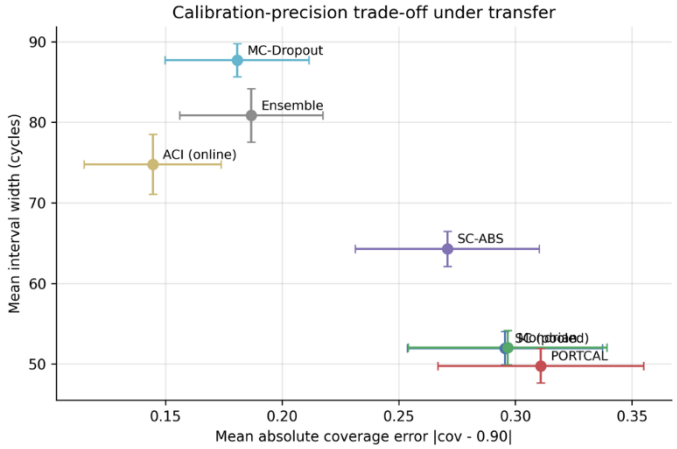


Fig. 4. Reported means of absolute coverage error and interval width, with reported standard errors over thirty-six blocks.

Note (*Fig. 4*): SC and Mondrian labels overlap in the original artwork; their table values are error 0.296 and 0.297, with widths 52.0 cycles for both. Blocks are dependent, so standard errors are descriptive summaries.

6.3 | RQ3: Online Adaptation

The clipped ACI replay has reported mean coverage 0.755, absolute coverage error 0.145 and Winkler score 221.1, approximately 36.4% below SC. Mean width is 74.8 cycles, approximately 43.8% above SC. Thus it improves the pooled endpoints while leaving a fourteen-and-a-half-percentage-point coverage shortfall. The reported harshest block reaches only 0.133 coverage, while a favorable BiGRU FD004→FD002 block reaches 0.894. These results cannot be explained by a slowly engaging unconditional ACI guarantee: the tested projection and finite quantile support differ from the unprojected algorithm [10]. *Fig. 5* illustrates selected retrospective streams, not a prospective RUL-feedback deployment.

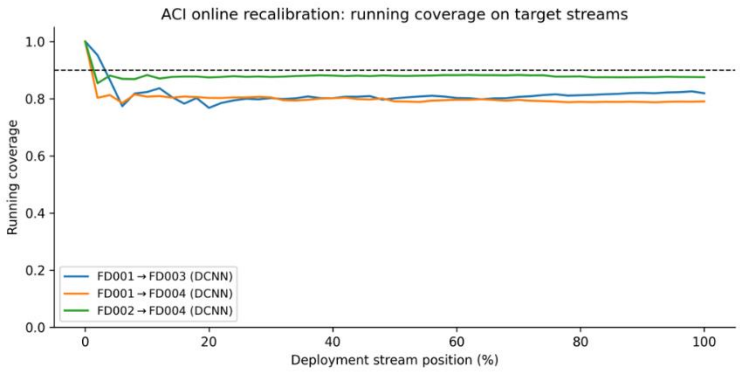


Fig. 5. Reported running coverage of clipped ACI on three DCNN transfer replays. Immediate realized RUL feedback is supplied after each interval; dashed line marks nominal 0.90. The displayed traces finish below nominal.

6.4 | Ablations

Replacing PORTCAL weights by equal weights reduces it to pooled SC and lowers the reported mean Winkler score by 10.5 cycles, with a separately reported paired p-value of 8.2×10^{-4}. Replacing normalized scores by absolute residuals lowers SC's mean Winkler score by about 37.0 cycles, with reported $p = 0.09$, and lowers absolute coverage error by 0.025. Mondrian differs from SC by +1.1 cycles, with $p = 0.91$. These are configuration contrasts, not equivalence tests or a diagnosis of the shift mechanism. The unadjusted ablation p-values must be distinguished from *Table 4's* adjusted family.

6.5 | Point Accuracy under Transfer

Table 5 reports final-observation point metrics across all sixteen tasks per backbone, including four in-domain and twelve transfer tasks. Mean RMSE is 45.8±17.7, 42.6±17.8 and 42.0±17.1 cycles for DCNN, BiGRU and TRANS; in-domain values are 27.5, 26.5 and 25.2. These metrics use seed-101 predictions clipped to [0,130], rather than demonstrating that every interval strategy has the same point center. For test unit i, the supplied final RUL anchors earlier labels by adding the number of cycles before its last observation. Point metrics and interval-window metrics use different evaluation units and should be interpreted separately.

Table 5. Point metrics at final observation with the standard [0,130] prediction clip (cycles; PHM08 score lower is better).

Backbone	RMSE (all 16 Tasks)	In-Domain RMSE	MAE	PHM08 Score
DCNN	45.8±17.7	27.5	37.1	5055
BIGRU	42.6±17.8	26.5	34.4	4265
TRANS	42.0±17.1	25.2	33.9	1922

6.6 | Per Task Records and Reproducibility

The study describes forty-eight per-task records with interval metrics, unit counts and diagnostics for weights, cluster occupancy, adapted levels and predictive variance. Those records would permit verification of aggregates, ranks and paired tests. An accessible archive is not identified in the manuscript; no assertion of a completed public release or bitwise programmatic verification is made here.

7 | Discussion

Weighted conformal theory applies under its weighted exchangeability and covariate-shift assumptions [9]. The observed PORTCAL result shows that this particular weighting procedure does not repair the reported benchmark coverage. It does not falsify the oracle theory or prove that label and concept shift dominate. Diagnosing the failure requires checking ratio quality, support overlap, target conditional residuals, normalization, clipping and trajectory dependence.

ACI replay and the uncalibrated baselines improve some transfer metrics but remain below nominal coverage. Their information contracts differ: ACI consumes retrospective outcome feedback, PORTCAL uses unlabeled target features, and ensembles train multiple members. A deployment comparison must account for label delay, training and inference costs, and maintenance loss. No tested option is certified for cross-fleet deployment by these summaries.

The empirical result supports reporting both in-domain and transfer coverage on public benchmarks. Prior conformal RUL research [7] supplies methods to compare, but its guarantees should be read under the original assumptions. The present experiment evaluates changed subsets and modified wrappers; the appropriate conclusion is a scope-specific coverage audit rather than a general dismissal of conformal RUL validity.

Normalized conformal intervals retain input dependence through $\sigma(X)$, even when q is constant on a calibration domain. Under shift, both the predicted scale and the calibrated multiplier can be inaccurate. A useful extension would compare normalized residual calibration with calibration of ensemble spread and

independently assessed out-of-distribution scale adjustment. Such variants require experiments and do not follow from the present interval means alone.

A useful reporting contract states nominal coverage, empirical coverage in each domain, mean width, Winkler score, evaluation unit and the timing of target information. Unit-level uncertainty and task-level variability should accompany window aggregates. This permits a reader to distinguish statistical calibration, retrospective replay and operational feasibility.

8 | Limitations and Threats to Validity

The study evaluates simulated C-MAPSS subsets. It does not cover real fleet heterogeneity or the bearing transfer families mentioned in the initial design. The reported coverage differences are neither a lower nor an upper bound on field failures. Model capacity, target clipping and selected sensor channels further restrict the comparison.

The backbones use eight fixed epochs and no architecture search; the ensemble has five members and MC-dropout twenty draws. Stronger predictors and alternative training losses could change residuals, coverage and ordering. The common source protocol controls some differences, but ensemble cost and feedback availability are not equal across strategies.

PORTCAL uses per-window means and clipped weights, while its finite threshold safeguard differs from the oracle procedure. The experiment reports that this safeguard never fired. ACI uses $\gamma = 0.02$, projection and fixed calibration scores; its poor-coverage blocks show that this implementation should not inherit the unprojected algorithm's guarantee. The replay provides immediate RUL labels that may be unavailable until a unit fails in practice.

Transfer blocks share trained sources and target trajectories, and windows within each unit overlap. Conventional independent-block tests and window-count uncertainty may therefore overstate precision. Aggregate ranks, paired signs and the Nemenyi normalization require original-record verification. Exact unit IDs, indexing, scaling, sensor selection, cluster fallbacks and interval endpoint conventions are necessary to distinguish procedural failures from benchmark effects.

Replication requires the experiment's archive checksum, preprocessing and unit manifests, sixty model checkpoints, seed settings, evaluation code, per-task outcomes and figure-generation scripts. The canonical raw FD004 archive contains 249 training units and 248 test units, consistent with *Table 1's* training/calibration total and test count [4]. Exact unit allocation and windowing still require the experiment's manifest. Code availability and author declarations remain confirmations rather than established release claims.

9 | Conclusion

The reported C-MAPSS experiment shows a substantial decrease in pooled SC coverage between in-domain and cross-subset tasks. The tested PORTCAL weighting procedure has lower aggregate coverage and higher mean Winkler score than SC. Clipped ACI with immediate retrospective feedback improves both metrics but remains below nominal, while dropout and ensemble intervals trade greater width for higher transfer coverage than offline conformal variants. These observations favor explicit domain-by-domain reporting and comparison of information requirements. They do not uniquely diagnose the shift, validate a modified conformal guarantee or establish prospective maintenance reliability. Verification of the raw analyses, delayed-feedback experiments and broader benchmark testing would strengthen the conclusions.

References

[1] Li, X., Ding, Q., & Sun, J. Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability engineering & system safety*, *172*, 1–11. https://doi.org/10.1016/j.ress.2017.11.021

[2] Zhang, Z., Song, W., & Li, Q. (2022). Dual-aspect self-attention based on transformer for remaining useful life prediction. *IEEE transactions on instrumentation and measurement*, *71*, 1–11. https://doi.org/10.1109/TIM.2022.3160561

[3] Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. *2008 international conference on prognostics and health management* (pp. 1–9). IEEE. https://doi.org/10.1109/PHM.2008.4711414

[4] NASA Prognostics Center of Excellence (PCoE). (2025). *CMAPSS Jet engine simulated data*. https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data

[5] Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International conference on machine learning* (pp. 1050–1059). Proceedings of Machine Learning Research (PMLR). https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text

[6] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*, 1–15. https://doi.org/10.48550/arXiv.1612.01474

[7] Javanmardi, A., & Hüllermeier, E. (2023). *Conformal prediction intervals for remaining useful lifetime estimation*. https://doi.org/10.48550/arXiv.2212.14612

[8] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and trends in machine learning*, *16*(4), 494–591. https://doi.org/10.1561/2200000101

[9] Tibshirani, R. J., Foygel Barber, R., Candes, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, *32*. https://doi.org/10.48550/arXiv.1904.06019

[10] Gibbs, I., & Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in neural information processing systems*, *34*, 1660–1672. https://doi.org/10.48550/arXiv.2106.00170

[11] Gibbs, I., & Candès, E. J. (2024). Conformal inference for online prediction with arbitrary distribution shifts. *Journal of machine learning research*, *25*(162), 1–36. https://jmlr.org/papers/volume25/22-1218/22-1218.pdf

[12] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*(Jan), 1–30. https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[13] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. https://doi.org/10.2307/4615733