PHYSICS-AUDITABLE DRYOUT RECOGNITION IN HELICAL TUBES: CONDITION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable dryout recognition from sparse wall thermocouples supports thermal-margin assessment in helical tubes. Existing control-based and raw-channel models can confound condition-dependent temperature levels with circumferential film-redistribution patterns. The main bottleneck is condition transfer in the small-sample regime: pressure and mass flux shift the common temperature level, four-side sampling aliases higher-order circumferential modes, and random splits can obscure group dependence. We introduce a physics-auditable hybrid world-model architecture for retrospective report-label recognition. Its primary lane uses an exact four-point harmonic transform to separate the common level from signed circumferential contrasts before fixed-threshold ExtraTrees classification. An anchored residual RSSM and a differentiable FE decoder provide auxiliary rollout and conditional stress-audit branches. On the $3.5 \to 7$ MPa pressure-held-out split, the primary lane reaches BA 0.929, compared with 0.786 for raw channels and 0.823 for controls-only ExtraTrees. The five-fold group-held-out BA is 0.957, compared with 0.916 for raw channels. The gain is specific to fixed-threshold BA and does not extend uniformly to AUROC; the focal pressure comparison is not statistically significant. Anchor ablation shows severe rollout drift, although the RSSM remains slightly worse than persistence, and the FE branch remains a conditional structural screen rather than a measured-stress validation. This architecture provides an auditable route from sparse measurements to condition-shifted recognition with explicit evidence boundaries.

1 INTRODUCTION

Reliable scientific representations should preserve target-relevant structure when operating conditions change. This requirement is especially difficult with few labeled experiments and sparse sensors. Helical-tube dryout provides a concrete example. Curvature redistributes the liquid film around the wall, whereas pressure shifts both the absolute temperature level and its circumferential contrast. A model that learns this baseline can therefore appear accurate within one condition and fail when the condition changes.

We study retrospective, run-level report-label recognition from the first 100 s of sparse wall-temperature measurements. Figure 1 shows the measurement geometry: 15 axial stations along a helical tube, four circumferential sides at each station, and six nominal controls. The output is the historical report label for each run. This scope is deliberate. It evaluates recognition under condition shift rather than online event detection or early warning. A report label records how an experiment was classified, whereas an independently reviewed onset time would support a different scientific question.

This geometry exposes the central failure mode. Dryout-relevant information is expected to appear as circumferential liquid-film redistribution and therefore as asymmetric wall temperatures, but pressure and mass flux also shift the overall temperature level. With only four circumferential sensors per station and few labeled runs, an unconstrained model can memorize absolute temperature or static controls, perform well under same-condition splits, and fail on a new pressure or experiment group. The technical challenge is to separate condition-driven common-mode temperature changes from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

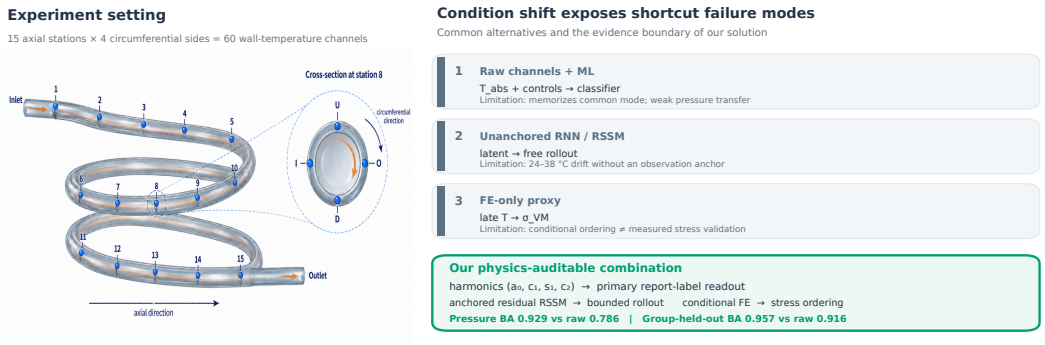


Figure 1: Experimental setting, failure modes, and our solution. Left: 15 axial stations and four circumferential sides provide 60 wall-temperature channels on the helical tube (schematic, not to scale). Right: common-mode shortcuts, unanchored rollout drift, and conditional FE assumptions motivate a separated architecture. The geometry-harmonic readout supplies the primary report-label evidence, an anchored residual RSSM stabilizes rollouts, and FE provides conditional stress ordering. Displayed balanced accuracies are retrospective results. The sensor convention is O=outer, U=upper, I=inner, and D=lower.

dryout-relevant circumferential structure while retaining a readout that can be audited against controls and physical rules.

We therefore use a modular, physics-auditable architecture. A fixed four-point harmonic projection separates the common level from inner–outer, upper–lower, and alternating contrasts. Contract-matched controls-only, raw-channel, outlet-quality, and group-held-out evaluations test whether any gain comes from sensor structure rather than shortcuts. An observation-anchored residual RSSM addresses free-rollout drift, and a conditional finite-element decoder maps late observed fields to a stress ordering under declared geometry and boundary assumptions. These branches expose distinct evidence types and are evaluated through their respective interfaces.

Our corpus contains 61 deduplicated historical runs, with 31 dryout and 30 non-dryout report labels, across two pressures and four mass-flux levels. We evaluate two complementary observation readouts. The geometry-harmonic readout is the primary pressure-transfer model. The contrast-covariance readout removes the mean, projects the three-dimensional zero-sum subspace with an orthonormal basis, and summarizes stationwise temporal covariance through spectral invariants. Both methods and their comparisons were developed retrospectively.

Our evaluation asks three questions. Does separating the common temperature level improve fixed-threshold report-label recognition under pressure transfer? Does recognition persist when entire experiment-label groups are held out, relative to quasi-static baselines? Which components of the representation explain the observed differences? We answer these questions with unified split reporting, paired uncertainty, and targeted representation ablations.

Contributions.

- **A modular physics-auditable world model.** A fixed geometry-harmonic observation lane supplies the primary report-label readout, while a physics-anchored residual RSSM and differentiable FE decoder provide explicit dynamics and stress-audit lanes with declared interfaces. The branches are co-designed and evaluated through separate evidence streams.
- **A condition-aware recognition evaluation.** Random, pressure-held-out, mass-flux-held-out, and experiment-label group-held-out results expose the role of quasi-static controls and outlet quality. Paired tests and bootstrap intervals distinguish observed effect sizes from evidence of superiority.
- **An auditable representation and stability analysis.** Transform ablations isolate common-mode separation and contrast fluctuations. The mean-free covariance extension rejects arbitrary additive common-mode drift for fixed controls, while anchor ablations quantify rollout stability and sensor-loss tests delineate the boundary of that guarantee.

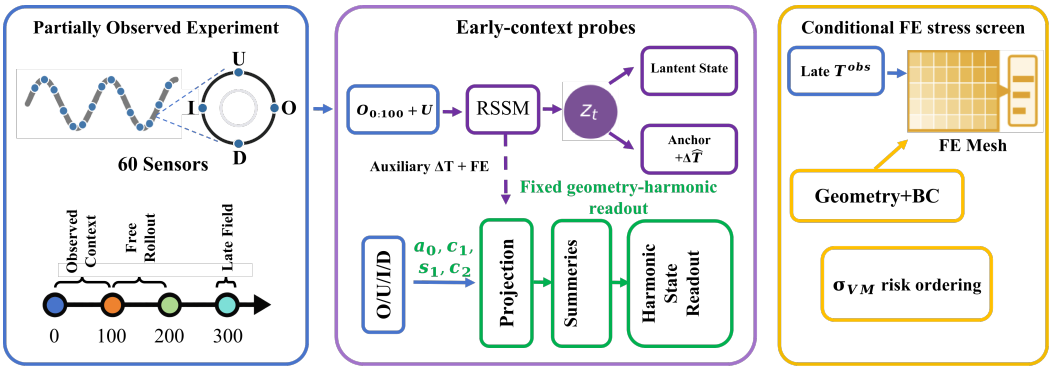


Figure 2: Physics-auditable world-model architecture. The left panel represents a partially observed experiment with 60 wall-temperature sensors and separates the observed context, free rollout, and late-field readouts. In the middle panel, early temperatures and controls feed an auxiliary RSSM probe with an explicit observation anchor, while the same early observations enter the fixed geometry-harmonic lane for the primary report-label readout. The right panel maps a late observed field, geometry, and boundary conditions to a conditional thermoelastic FE mesh and σ_{VM} risk ordering. Solid arrows denote branch-local inference, and the dashed arrow denotes auxiliary supervision. The contrast-covariance extension uses the same early observations, as defined in Section 2.2.

2 Physics-Auditable Observation Readouts

2.1 Fixed Circumferential Harmonics

Let $o_t \in \mathbb{R}^{60}$ denote the wall temperatures and $u_t \in \mathbb{R}^6$ the nominal controls (pressure, mass flux, heat flux, inlet quality, and two pressure-normalized dimensionless numbers $-\log_{10}\mathrm{Fr}_c$ and $\log_{10}\mathrm{Bo}$ computed with IAPWS-IF97 saturation states (Wagner et al., 2000)). At each of the 15 stations the four sides (O, U, I, D), sampled at circumferential angles $0, \pi/2, \pi, 3\pi/2$, are projected as

$$a_0 = \tfrac{O+U+I+D}{4}, \quad c_1 = \tfrac{O-I}{2}, \quad s_1 = \tfrac{U-D}{2}, \quad c_2 = \tfrac{O+I-U-D}{4}. \tag{1}$$

The projection is an invertible discrete Fourier representation of the four samples. It isolates the common temperature level in a_0 and expresses side-to-side asymmetry through c_1, s_1, c_2. Under independent per-channel noise with variance σ^2, the mode variances are $\sigma^2/4, \sigma^2/2, \sigma^2/2, \sigma^2/4$. The four-sensor layout observes a fixed-phase second harmonic, with continuous-field aliasing characterized in Appendix C. The contrast directions correspond to inner–outer asymmetry, upper–lower stratification, and an alternating circumferential pattern, respectively.

Each mode is summarized over the 0–100 s context by its mean, standard deviation, final value, full-window slope, and recent 20 s slope. Appending six controls gives 306 features. ExtraTrees uses 1000 trees, Gini splits, `min_samples_leaf` $= 2$, `max_features` $= \sqrt{d}$, balanced class weights, no bootstrap, and a threshold of 0.5. Raw-channel and controls-only baselines share the tree settings and splits. Linear, PCA, gradient-boosting, and random-feature comparisons test readout dependence. All settings are fixed for the reported evaluations.

2.2 Mean-Free Contrast Covariance Extension

For a station with temperature matrix $X_j \in \mathbb{R}^{T\times 4}$, we use the fixed orthonormal zero-sum basis

$$Q = \mathrm{diag}(1/\sqrt{2}, 1/\sqrt{2}, 1/2) \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \qquad Q\mathbf{1} = 0, \quad QQ^{\mathsf{T}} = I_3. \tag{2}$$

Write $Y_j = X_j Q^{\mathsf{T}}$ for the uncentered contrasts and $H = I_T - \mathbf{1}_T\mathbf{1}_T^{\mathsf{T}}/T$ for temporal centering. We form $Z_j = HY_j$ and $C_j = Z_j^{\mathsf{T}} Z_j/(T-1) + 10^{-6} I_3$. The descriptor concatenates six controls, the five temporal summaries of each coordinate of Y_j, and $\mathrm{tr}(C_j)$, $\mathrm{tr}(C_j^2)$, and $\log\det(C_j)$ at each station, giving $6 + 15\times 3\times 5 + 15\times 3 = 276$ features. Summarizing Y_j retains stationary signed

contrasts, while centering only the covariance isolates their temporal variation. Temperatures use degrees Celsius and the ridge uses the corresponding squared units.

Exact properties and their scope. For fixed controls, the entire descriptor is unchanged by $X_j \mapsto X_j + b_j \mathbf{1}_4^\top$ for any time-varying $b_j \in \mathbb{R}^T$, because $Q\mathbf{1}_4 = 0$. Under $Q \mapsto RQ$ with orthogonal R, only the covariance spectral block is invariant. The signed summaries and the complete classifier depend on the chosen basis. Appendix T gives the derivation. These are established projection and covariance properties (Tuzel et al., 2006, Barachant et al., 2012, Mroueh et al., 2015). Their use here specifies which measurement perturbations the representation removes exactly. The stated guarantee covers additive offsets; side gains and missing sensors require separate analysis. The classifier and split contract match the harmonic readout, but the feature count and retained information differ.

2.3 Physics-anchored residual world model

The recurrent lane is a dynamics probe for residual forecasting. Let T_t be the last observed wall-temperature field, q_t a low-rank residual state, h_t the deterministic recurrent state, and u_t the controls. The residual update is

$$q_{t+1} = \rho_t \odot q_t + (1 - \rho_t) \odot f_\theta(h_t, z_t, u_t), \qquad \widehat{T}_{t+1} = T_t + B q_{t+1}, \tag{3}$$

where B is a fixed modal basis and the observation anchor T_t is carried explicitly into every predicted field. The gate ρ_t is learned from the RSSM state, while z_t uses the observation-conditioned posterior during training and the prior during rollout. Removing the anchor produces the stability ablation reported in Table 5.

The predicted field is passed through a differentiable fixed FE temperature-to-stress operator,

$$\widehat{\sigma}_t = \mathcal{F}_{\mathrm{FE}}(\widehat{T}_t, \text{ geometry}, \text{boundary conditions}), \tag{4}$$

and the training objective combines one-step temperature error, report-label and mode losses, KL regularization, and a small FE stress-consistency term. The operator is a conditional structural decoder whose stress output is interpreted as a relative ordering under the specified geometry and boundary conditions. The RSSM and FE lanes therefore provide complementary dynamics and structural audits alongside the primary report-label readout.

3 Evaluation protocol

Content-hash deduplication gives 61 runs, with 31 dryout and 30 non-dryout report labels, nested in 48 experiment-label groups over the 2×4 pressure–mass-flux factorial of Table 1 (observation and split details in Appendix B). We evaluate four split families. Random run-level partitions use three fixed seeds, with a 20-seed extension in Appendix E. The pressure-held-out split is a single $3.5 \to 7$ MPa extrapolation contrast with $n_{\text{test}} = 33$. The mass-flux-held-out split has $n_{\text{test}} = 15$. Stratified group-held-out evaluation uses five folds over the 48 experiment-label groups, with a mean test size of 12.2 runs per fold. Normalization and model fitting use the prescribed training members. Each fitted model scores held-out members without adapting to their labels. All analyses are retrospective and use the same historical corpus.

Endpoint and threshold protocol. The primary endpoint is the mean BA difference between the geometry-harmonic readout and contract-matched controls-only ExtraTrees on the pressure-held-out split. This focal contrast organizes the retrospective analysis. All classifiers use a frozen probability threshold of 0.5, with no threshold adjustment on held-out labels. Tabular preprocessing and model fitting use train plus validation members. The label-free physical comparator uses $x_{\mathrm{out}} \geq 1$ and requires no fitted threshold. AUROC evaluates score ranking separately. The remaining model, split, metric, FE, RSSM, and generated-data comparisons are exploratory. A sensitivity analysis fits on train members alone and selects the threshold on validation members (Appendix D).

Uncertainty and paired inference. Table 2 uses 5,000 stratified bootstrap draws for descriptive uncertainty. Standard splits resample unique run IDs within label strata, sharing each run's multiplicity

Table 1: Operating-condition envelope of the 61 historical runs (water/steam, heated helical tube, 15 axial stations × four circumferential sides).

Pressure (MPa)	Mass flux ($\mathrm{kg\,m^{-2}s^{-1}}$)	Heat flux ($\mathrm{kW\,m^{-2}}$)	Inlet quality (–)	Runs	Dryout
3.5	320	50–150	0.30–0.90	7	3
3.5	530	100–300	0.30–0.90	7	3
3.5	800	150–450	0.30–0.95	7	3
3.5	1100	200–600	0.30–0.95	7	3
7.0	320	50–150	0.40–0.95	8	4
7.0	530	50–300	0.30–0.95	9	6
7.0	800	75–450	0.30–0.95	8	5
7.0	1100	120–600	0.30–0.97	8	4

Table 2: Core retrospective report-label recognition. Entries are mean [95% bootstrap interval]. BA and AUROC are reported separately. Panel A averages three fixed fits. Panel B averages five experiment-label group-held-out folds covering all 61 runs in 48 groups. Panel A reports test-run counts and negative/positive labels. Random-split class counts are listed in seed order with vertical bars.

Model	Metric	Random	Pressure	Mass flux
	n_{test} (neg./pos.)	10 (3/7 | 5/5 | 5/5)	33 (14/19)	15 (8/7)
A. Fixed run-level splits				
Harmonic ExtraTrees	BA	0.967 [0.889, 1.000]	0.929 [0.821, 1.000]	0.938 [0.812, 1.000]
	AUROC	0.960 [0.850, 1.000]	0.917 [0.782, 1.000]	0.905 [0.679, 1.000]
Raw-channel ExtraTrees	BA	0.967 [0.889, 1.000]	0.786 [0.643, 0.929]	0.875 [0.688, 1.000]
	AUROC	1.000 [1.000, 1.000]	0.901 [0.756, 1.000]	0.917 [0.714, 1.000]
Controls-only ExtraTrees	BA	0.616 [0.405, 0.817]	0.823 [0.689, 0.939]	0.938 [0.812, 1.000]
	AUROC	0.615 [0.325, 0.844]	0.896 [0.747, 0.992]	0.893 [0.625, 1.000]
Controls-only logistic	BA	0.806 [0.650, 0.958]	0.893 [0.786, 1.000]	0.812 [0.625, 0.938]
	AUROC	0.897 [0.750, 1.000]	0.902 [0.748, 1.000]	0.839 [0.589, 1.000]
Outlet quality $x_{\mathrm{out}} \geq 1$	BA	0.910 [0.772, 1.000]	0.885 [0.771, 0.974]	0.652 [0.446, 0.857]
	AUROC	1.000 [1.000, 1.000]	0.929 [0.786, 1.000]	0.875 [0.625, 1.000]
Contrast-covariance ExtraTrees	BA	0.967 [0.889, 1.000]	0.876 [0.752, 0.974]	0.938 [0.812, 1.000]
	AUROC	0.960 [0.850, 1.000]	0.896 [0.742, 1.000]	0.881 [0.625, 1.000]
Dispersion ExtraTrees	BA	0.967 [0.889, 1.000]	0.860 [0.741, 0.964]	0.938 [0.812, 1.000]
	AUROC	1.000 [1.000, 1.000]	0.920 [0.774, 1.000]	0.893 [0.625, 1.000]

B. Group-held-out sensitivity (61 runs, 30/31 labels)

Model	BA	AUROC	Test runs per fold
Harmonic ExtraTrees	0.957 [0.908, 1.000]	0.984 [0.940, 1.000]	9, 13, 14, 14, 11
Raw-channel ExtraTrees	0.916 [0.856, 0.971]	0.987 [0.955, 1.000]	9, 13, 14, 14, 11
Controls-only ExtraTrees	0.586 [0.458, 0.710]	0.686 [0.554, 0.825]	9, 13, 14, 14, 11
Outlet quality $x_{\mathrm{out}} \geq 1$	0.905 [0.846, 0.961]	0.971 [0.911, 1.000]	9, 13, 14, 14, 11

BA uses threshold 0.5 for all fitted classifiers and $x_{\mathrm{out}} \geq 1$ for the label-free physical rule. AUROC measures score ranking, not probability calibration. Panel A resamples unique runs within label strata with shared weights across seeds. Panel B resamples experiment-label groups within folds. Intervals condition on fixed models and split assignments, so they exclude training and model-selection uncertainty. An AUROC interval of $[1, 1]$ reflects perfect ordering in the observed sample, not certainty about population performance. The retrospectively designated primary pressure contrast has exact McNemar $p = 0.125$, 0.125, and 0.250 across seeds. It is not significant at 0.05. Panel B includes models with available grouped predictions and the deterministic physical rule.

across all fixed seeds and models. Group-held-out evaluation resamples experiment-label groups within each test fold. The intervals condition on fitted models and split assignments. They do not include retraining or model-selection uncertainty. Exact two-sided McNemar tests assess paired correctness through discordant run decisions, not equality of BA or AUROC, and are reported separately from the bootstrap intervals. Seeds reuse experimental units and their p values are not combined. The additional permutation, cell-cluster, and calibration analyses are specified in Appendix D.

4 Results

4.1 Condition-transfer recognition

Pressure transfer shows a higher observed BA operating point. Harmonic ExtraTrees reaches BA 0.929, compared with 0.823 for contract-matched controls-only ExtraTrees and 0.786 for raw

Table 3: Unified paired audit across all three split families. ΔBA is the candidate-minus-baseline mean over three seeds (the corresponding BA/AUROC point estimates are in Table 2). W/T/L is candidate wins/ties/losses in held-out run correctness, shown per seed (identical triplets are compressed as ×3). The interval is the paired run-bootstrap 95% CI for the correctness difference, p is the two-sided exact McNemar test on discordant runs. The auxiliary GBM and random-feature rows in Table 2 are retained as point-estimate readout probes because their frozen artifact does not contain per-run correctness vectors.

Split	Baseline	ΔBA	W/T/L	Paired bootstrap 95% CI	Exact p
Random ($n = 10$)	Raw-channel ExtraTrees	0.000	0/10/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Random ($n = 10$)	Controls-only ExtraTrees	0.351	5/5/0, 3/7/0, 3/7/0	[0.200, 0.800], [0.000, 0.600], [0.000, 0.600]	0.062, 0.250, 0.250
Random ($n = 10$)	Controls-only logistic	0.160	4/6/0, 0/10/0, 1/9/0	[0.100, 0.700], [0.000, 0.000], [0.000, 0.300]	0.125, 1.000, 1.000
Random ($n = 10$)	Contrast-covariance ExtraTrees	0.000	0/10/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Random ($n = 10$)	Generic dispersion ExtraTrees	0.000	0/10/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Random ($n = 10$)	$x_{\mathrm{out}} \geq 1$ rule	0.057	1/9/0, 1/9/0, 1/8/1	[0.000, 0.300], [0.000, 0.300], [-0.300, 0.300]	1.000 (×3)
Pressure ($n = 33$)	Raw-channel ExtraTrees	0.143	4/29/0 (×3)	[0.030, 0.242] (×3)	0.125 (×3)
Pressure ($n = 33$)	Controls-only ExtraTrees	0.106	4/29/0, 4/29/0, 3/30/0	[0.030, 0.242], [0.030, 0.242], [0.000, 0.212]	0.125, 0.125, 0.250
Pressure ($n = 33$)	Controls-only logistic	0.036	1/32/0 (×3)	[0.000, 0.091] (×3)	1.000 (×3)
Pressure ($n = 33$)	Contrast-covariance ExtraTrees	0.053	2/31/0 (×3)	[0.000, 0.152] (×3)	0.500 (×3)
Pressure ($n = 33$)	Generic dispersion ExtraTrees	0.068	1/32/0, 3/30/0, 2/31/0	[0.000, 0.091], [0.000, 0.212], [0.000, 0.152]	1.000, 0.250, 0.500
Pressure ($n = 33$)	$x_{\mathrm{out}} \geq 1$ rule	0.043	3/29/1 (×3)	[-0.061, 0.182] (×3)	0.625 (×3)
Mass flux ($n = 15$)	Raw-channel ExtraTrees	0.062	1/14/0 (×3)	[0.000, 0.200] (×3)	1.000 (×3)
Mass flux ($n = 15$)	Controls-only ExtraTrees	0.000	0/15/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Mass flux ($n = 15$)	Controls-only logistic	0.125	2/13/0 (×3)	[0.000, 0.333] (×3)	0.500 (×3)
Mass flux ($n = 15$)	Contrast-covariance ExtraTrees	0.000	0/15/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Mass flux ($n = 15$)	Generic dispersion ExtraTrees	0.000	0/15/0 (×3)	[0.000, 0.000] (×3)	1.000 (×3)
Mass flux ($n = 15$)	$x_{\mathrm{out}} \geq 1$ rule	0.286	4/11/0 (×3)	[0.067, 0.533], [0.067, 0.468], [0.067, 0.533]	0.125 (×3)

channels (Table 2). The corresponding ΔBA values are $+0.106$ and $+0.143$. Against controls, the candidate corrects four runs in two seeds and three in the third, with no losses. Exact McNemar p values are 0.125, 0.125, and 0.250, so the observed effect is not statistically significant at 0.05 (Table 3). The advantage is principally fixed-threshold BA. Harmonic AUROC is 0.917 versus 0.901 for raw channels on pressure transfer, but is lower than raw AUROC on both random and mass-flux splits.

Quasi-static baselines explain much of the task. Controls-only logistic regression reaches pressure BA 0.893, just 0.036 below harmonic ExtraTrees. The zero-parameter outlet-quality rule reaches BA 0.885 and AUROC 0.929, exceeding harmonic AUROC on this split. On mass-flux transfer, controls-only ExtraTrees ties harmonic BA at 0.938, while the physical rule drops to 0.652. These contrasting outcomes make controls-only and outlet-quality models essential comparators for separating representation gains from operating-condition information. A contract-matched harmonic logistic audit reaches pressure BA 0.788 and mass-flux BA 0.938, with AUROC 0.876 and 0.875, respectively. The gap between this linear readout and harmonic ExtraTrees makes the headline gain a representation–readout interaction rather than evidence that the fixed harmonic coordinates dominate every classifier. Full three-seed linear results are reported in Appendix E.

Group-held-out recognition remains strong. Holding out complete experiment-label groups gives harmonic BA 0.957, compared with 0.916 for raw channels, 0.586 for controls-only ExtraTrees, and 0.905 for outlet quality. These are means over five folds, with all 61 runs scored out of fold. Harmonic AUROC is 0.984, compared with 0.987 for raw channels and 0.971 for outlet quality. Group separation therefore preserves strong thresholded recognition, while AUROC remains similar to raw measurements. Gradient-boosting and random-feature readout comparisons remain in Appendix E, with their point-estimate artifact coverage reported there.

Common-mode separation and contrast fluctuations account for the pressure-transfer behavior. Demeaning each station's four channels reaches 0.876, while preserving the mean and randomly rotating the three contrast directions reaches 0.914 ± 0.031 over 20 draws. Contrast standard deviations alone recover 0.929, compared with 0.875 for the linear summaries together. These results identify a useful representation principle: separate the common temperature level before summarizing circumferential fluctuations. A dimension-matched generic baseline using station mean, spatial standard deviation, range, and mean absolute deviation reaches 0.860 ± 0.031 on pressure and 0.938 on mass flux. Its pressure AUROC is 0.920, comparable to the harmonic 0.917. Thus side-invariant dispersion recovers part of the thresholded gain, while the signed harmonic coordinates retain an advantage for this ExtraTrees operating point. The fixed harmonic basis supplies common-mode separation

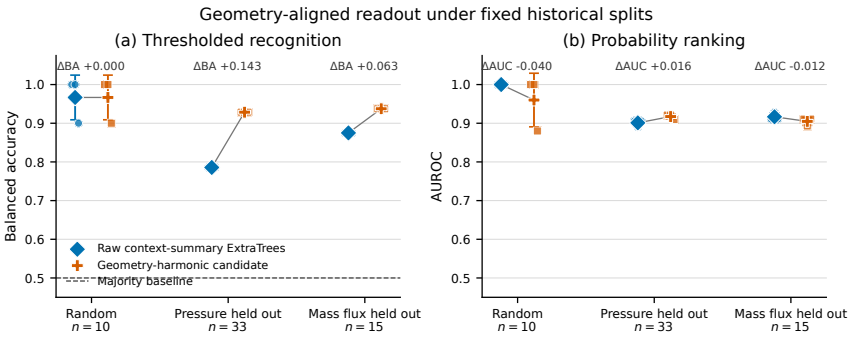


Figure 3: Geometry-harmonic and raw-channel report-label recognition across the fixed run-level splits. Points and error bars show fixed-fit means and sample SD. Labels show candidate-minus-raw effects at the frozen threshold. Group-held-out results are reported in Table 2.

with physically interpretable coordinates. Full transform nulls and condition-preserving permutations appear in Appendix I.

What the ablations identify. The projection is invertible, so it adds no measurements. Its effect comes from presenting common levels and contrasts separately to temporal summaries and axis-aligned tree splits. Standard deviations do not commute with channel mixing, which changes the fluctuation features even when the pointwise transform preserves information. The rotation and matched-readout results therefore support a representation–classifier interaction rather than a uniquely necessary harmonic basis. The covariance extension also removes the mean and changes feature dimension, so its results describe the full extension rather than the spectral block alone.

The benefit varies with the transfer setting. The second harmonic alone recovers the full pressure and mass-flux results, while reversing pressure transfer ties harmonic and raw at 0.917. Group-held-out BA is 0.957 for harmonic, 0.916 for raw, and 0.586 for controls, with AUROC 0.984, 0.987, and 0.686. Cell-held-out covariance has the highest pooled BA point estimate, but its one-run per-seed margin over harmonic is not significant ($p = 1.0$). These results localize the largest observed difference to the low-to-high pressure shift rather than to universal dominance.

Sensor perturbations expose a failure boundary. With the fitted model and controls held fixed, a $2\,^{\circ}\mathrm{C}$ common-mode perturbation leaves contrast-covariance probabilities unchanged. Relative side gains give BA 0.911, while a missing bottom side lowers it from 0.876 to 0.753. The representation is invariant to common-mode offsets but remains sensitive to missing sensors. Appendix T.1 reports all splits.

4.2 Mechanism-consistent association audit

The circumferential contrasts retain label-associated structure within operating conditions (Figure 4). Cell-balanced dryout-minus-nondryout differences are $-8.00\,^{\circ}\mathrm{C}$ for c_1 and $+1.65\,^{\circ}\mathrm{C}$ for c_2, with the inner-hotter direction agreeing in all eight cells. Within-cell permutations give Holm-adjusted $p = 0.0004$ for each mode (Appendix S). This association is consistent with the proposed film-redistribution mechanism and links the report label to the measured circumferential structure.

4.3 Mechanism, identifiability, and failure boundaries

Three audits connect transfer to the proposed mechanism (Figure 5). The mean shifts strongly across pressure ($|\mathrm{SMD}| = 1.70$), while the contrasts shift less. The harmonic readout retains the highest mean BA in repeated small-budget fits (0.793–0.920). Generated mechanism checks expose aliasing and condition-shift boundaries; prospective data are required to evaluate generalization beyond the historical corpus.

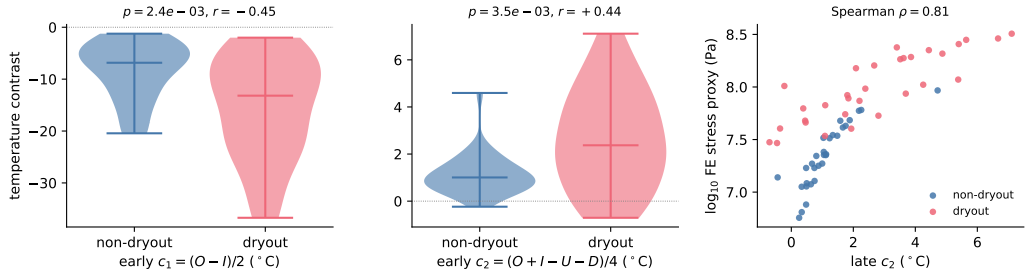


Figure 4: Physics audit on all 61 runs. Label-conditional distributions of the early context (0–100 s) station-mean contrasts $c_1 = (O - I)/2$ and $c_2 = (O + I - U - D)/4$ (two-sided Mann–Whitney p and rank-biserial effect r), and the late (290–310 s) second harmonic against the finite-element stress proxy. Dryout runs carry a roughly doubled inner-hotter c_1 and elevated c_2.

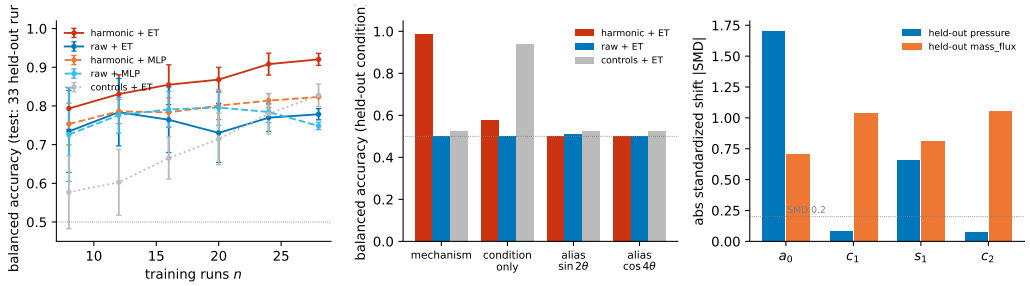


Figure 5: Geometry-prior audit. (a) Sample-efficiency on the real leave-pressure axis. (b) A development-generated check reproduces the designed mechanism and condition regimes and exposes sensor-layout failure cases. (c) Mode-level domain shift on the real corpus. The generated panels exercise the frozen interfaces and illustrate their operating boundaries.

4.4 Auxiliary thermoelastic consistency check

The FE branch is secondary to the report-label endpoint. On the historical late fields, its conditional stress proxy reaches ROC-AUC 0.929 with within-cell micro-AUC 0.973, while a manufactured-solution suite verifies the Hex8 assembly with a stress error of 6.56×10^{-7} Pa. These checks support numerical consistency and conditional stress ordering under the stated assumptions. Absolute stress calibration requires measured mechanical data. Full resolution and boundary-condition audits are in Appendix L.

4.5 Auxiliary recurrent probes and forecasting boundary

We evaluate recurrent probes as auxiliary dynamics models. GRU and RSSM latent states retain pressure-axis ranking performance near AUROC 0.94, but their fixed-threshold BA depends strongly on the downstream readout. RSSM prior rollouts have median MAE 32.88 °C on the pressure split versus 0.233 °C for persistence. A physics-anchored residual RSSM with a differentiable FE stress decoder retains the observed field as an explicit anchor and prevents catastrophic drift. Removing the anchor increases temperature MAE to 24–38 °C across split families, while the anchored model remains slightly worse than persistence in temperature MAE (Table 5). Persistence is therefore retained as the stationary forecast comparator; report-label recognition is evaluated separately in Table 2. Complete matched-readout and physics-anchored results are in Appendix M.

5 Related work

Helical-tube heat transfer is governed by curvature-induced secondary flow and circumferential film redistribution (Dean, 1927; 1928, Berger et al., 1983). Dryout and critical-heat-flux studies report

nonuniform film behavior and asymmetric wall temperatures (Kaji et al., 1995, Santini et al., 2014, Hwang et al., 2014, Chung et al., 2014, Xu et al., 2022), while recent CHF models mainly learn from operating-condition databases (Chen et al., 2011, Yang et al., 2024, Pedroni, 2025, Wang et al., 2026). Our setting adds a sparse spatial observation problem: the model must use four circumferential measurements without treating the pressure-dependent baseline as a stable physical signature.

The covariance branch builds on established second-order descriptors and group-invariant representations (Tuzel et al., 2006, Barachant et al., 2012, Mroueh et al., 2015). We instantiate these ideas with a fixed zero-sum basis, temporal covariance spectra, and declared sensor perturbations. Scientific machine-learning methods such as neural operators, DMDc, and physics-informed learning target field-to-field maps and controlled dynamics (Li et al., 2021, Lu et al., 2021, Proctor et al., 2016, Raissi et al., 2019, Karniadakis et al., 2021). Our modular composition uses the geometry-harmonic lane for recognition, the anchored RSSM for residual rollouts, and the FE decoder for conditional structural screening. The branches provide complementary evidence types and are evaluated separately.

6 Discussion and conclusion

The architecture combines a geometry-aware readout, an anchored residual dynamics lane, and a conditional FE stress decoder. The harmonic lane supplies recognition evidence, the anchor ablation establishes rollout stability, and the FE lane supplies structural ordering under stated assumptions. The primary pressure comparison is retrospective and its paired test is not statistically significant. The results nevertheless establish an auditable branch interface for condition-shifted recognition. Independent acquisitions, reviewed event times, traceable calibration, and measured stress are the next steps for extending the evidence.

Reproducibility statement

Results are linked to internally frozen artifacts and SHA256 checksums. Raw measurements and source code are withheld under data-ownership and intellectual-property restrictions; the retained bundle supports internal audit.

Ethics statement

The study uses non-human, non-animal thermal experiments.

References

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012. doi: 10.1109/TBME.2011.2172210.

S. A. Berger, L. Talbot, and L.-S. Yao. Flow in curved pipes. *Annual Review of Fluid Mechanics*, 15: 461–512, 1983. doi: 10.1146/annurev.fl.15.010183.002333.

C. N. Chen, J. T. Han, T. C. Jen, and L. Shao. Dry-out chf correlation for r134a flow boiling in a horizontal helically-coiled tube. *International Journal of Heat and Mass Transfer*, 54(1–3): 739–745, 2011. doi: 10.1016/j.ijheatmasstransfer.2010.09.027.

Young-Jong Chung, Kyoo-Hwan Bae, Keung Koo Kim, and Won-Jae Lee. Boiling heat transfer and dryout in helically coiled tubes under different pressure conditions. *Annals of Nuclear Energy*, 71: 298–303, 2014. doi: 10.1016/j.anucene.2014.04.015.

W. R. Dean. Xvi. note on the motion of fluid in a curved pipe. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(20):208–223, 1927. doi: 10.1080/14786440708564324.

W. R. Dean. Fluid motion in a curved channel. *Proceedings of the Royal Society of London. Series A*, 121(787):402–420, 1928. doi: 10.1098/rspa.1928.0205.

Kyung Won Hwang, Dong Eok Kim, Ki Hoon Yang, Jin Man Kim, Moo Hwan Kim, and Hyun Sun Park. Experimental study of flow boiling heat transfer and dryout characteristics at low mass flux in helically-coiled tubes. *Nuclear Engineering and Design*, 273:529–541, 2014. doi: 10.1016/j.nucengdes.2014.03.046.

Masou Kaji, Koji Mori, Shigeyasu Nakanishi, Kentaro Hirabayasi, and Masaya Ohishi. Dryout and wall temperature fluctuations in helically coiled evaporating tubes. *Transactions of the Japan Society of Mechanical Engineers Series B*, 61(585):1811–1817, 1995. doi: 10.1299/kikaib.61.1811.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.08895.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Youssef Mroueh, Stephen Voinea, and Tomaso A. Poggio. Learning with group invariant features: A kernel perspective. In *Advances in Neural Information Processing Systems 28*, 2015.

Nicola Pedroni. Ensembles of physics-enhanced neural networks for the prediction of critical heat flux in nuclear reactors and the quantification of its uncertainty. *Frontiers in Nuclear Engineering*, 4:1692182, 2025. doi: 10.3389/fnuen.2025.1692182.

Joshua L. Proctor, Steven L. Brunton, and J. Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016. doi: 10.1137/15M1013857.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Lorenzo Santini, Andrea Cioncolini, Carlo Lombardi, and Marco Ricotti. Dryout occurrence in a helically coiled steam generator for nuclear power application. *EPJ Web of Conferences*, 67:02102, 2014. doi: 10.1051/epjconf/20146702102. Experimental helical steam-generator dryout study over varied pressure, mass flux, and heat flux conditions.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pp. 589–600, 2006. doi: 10.1007/11744023_46.

W. Wagner, J. R. Cooper, A. Dittmann, J. Kijima, H.-J. Kretzschmar, A. Kruse, R. Mares, K. Oguchi, H. Sato, I. Stoecker, O. Sifner, Y. Takaishi, I. Tanishita, J. Truebenbach, and Th. Willkommen. The iapws industrial formulation 1997 for the thermodynamic properties of water and steam. *Journal of Engineering for Gas Turbines and Power*, 122(1):150–182, 2000. doi: 10.1115/1.483186.

Ningyuan Wang, Deqi Chen, Hanzhou Liu, Shanshan Bu, and Hongxing Yu. Prediction model for the initial dryout vapor quality in helically coiled tubes. *Energy*, 347:140377, 2026. doi: 10.1016/j.energy.2026.140377. Mechanistic and experimental model based on annular-film distribution, secondary flow, and interfacial shear.

Ziyi Xu, Maolong Liu, Cong Chen, Yao Xiao, and Hanyang Gu. Development of an analytical model for the dryout characteristic in helically coiled tubes. *International Journal of Heat and Mass Transfer*, 186:122423, 2022. doi: 10.1016/j.ijheatmasstransfer.2021.122423.

Kuang Yang, Haifan Liao, Bo Xu, Qiuxiang Chen, Zhenghui Hou, and Haijun Wang. Data-driven dryout prediction in helical-coiled once-through steam generator: A physics-informed approach leveraging the buckingham pi theorem. *Energy*, 294:130822, 2024. doi: 10.1016/j.energy.2024.130822. DimNet uses Buckingham-Pi dimensionless groups and noise augmentation for helical-coil dryout-quality prediction.

AI USE STATEMENT

The authors used large language models and other AI tools to assist with writing and language polishing, retrieval and discovery of related work, research ideation and execution, drafting sections, and checking mathematical claims. The authors independently reviewed the manuscript, verified the mathematical statements and reported results, and take responsibility for the final submission. No AI tool was used to generate synthetic datasets for the reported experiments.

A SCOPE, ASSUMPTIONS, AND EVIDENCE BOUNDARIES

This supplement is the anonymous protocol and evidence ledger for the accompanying paper. All evidence in the main text is retrospective and internal to one historical corpus containing 61 deduplicated experimental runs, 31 report-level dryout labels, and 30 non-dryout labels. Each report label annotates a run rather than an independently reviewed onset time. All audits reuse these 61 runs, and the report labels, late observed fields, and finite-element (FE) artifacts share historical provenance. The leave-pressure axis is a single $3.5 \to 7$ MPa contrast because the corpus contains exactly two pressure levels. Balanced accuracy of the frozen majority-class baseline is 0.5. Raw measurements cannot be publicly released because of data-ownership restrictions, and source code cannot be publicly released because of intellectual-property restrictions. Internally frozen prediction tables, split manifests, and checksums are retained for audit. The probability table contains genuine predictions with checked labels, splits, and recomputed metrics for models with per-run artifacts; source workbooks remain unavailable.

AI4S evidence ladder and intended use. The paper maps each claim to its supporting evidence. Historical recognition measures report-label transfer on fixed retrospective splits; cell-conditioned association tests persistence of the thermal contrast after coarse pressure–mass-flux stratification; and sensor perturbations quantify stability under declared nuisance models. Manufactured thermoelastic solutions verify the numerical kernel, while late measured fields provide a conditional structural ordering under specified material, geometry, and support conditions. Prospective event recognition, causal dryout mechanisms, absolute stress, fatigue, and deployment safety require additional evidence.

B OPERATING CONDITIONS AND OBSERVATION CONTRACT

The corpus is a 2×4 factorial of operating cells. Pressures are $\{3.5, 7.0\}$ MPa and mass fluxes are $\{320, 530, 800, 1100\}$ kg m^{-2}s^{-1}. Heat fluxes range from 50–600 kW m^{-2} and inlet qualities range from 0.30–0.97 within cells. The working fluid is water/steam, with IAPWS-IF97 saturation states used for Fr_c and Bo. All runs use one helical-tube geometry with 14 turns, a 0.112 m helix diameter, a 0.0225 m pitch, and 6/9 mm inner/outer radius. Cell membership is 7–9 runs per cell with 3–6 dryout labels each. The full table is frozen in `conditions_and_rollout_context.json`.

Each run records 60 outer-wall temperature channels: 15 axial stations crossed with four circumferential sides (outer, upper, inner, lower), plus six operating controls held at their nominal set points. The fixed audit windows are disjoint. Recognition uses the first 100 s. Forecast evaluation uses 100–200 s. The structural screen uses the observed field averaged over 290–310 s. These windows are fixed by the audit contract. The 200–290 s gap is not modeled. Content-hash deduplication precedes all splitting. Run identifiers are unique within the analysis.

SPLIT CONTRACT

The main table uses three fixed-fit split families with seeds 20260909, 20260910, and 20260911. These are random run-level partitions with $n_{\text{test}} = 10$, the highest pressure level held out with $n_{\text{test}} = 33$, and the highest mass-flux level held out with $n_{\text{test}} = 15$. The latter two provide single-contrast extrapolation checks rather than full-factorial coverage. The test set is deterministic and identical across seeds for these two families, so seed variation measures training-side randomness only. A separate five-fold StratifiedGroupKFold audit holds out all runs in each experiment-label group. It covers 48 groups and has a mean test size of 12.2 runs per fold. Tabular readouts fit on train and validation members. Test members are scored once at the frozen threshold of 0.5. The

GRU fits normalization on train members only. A 20-seed extension of the random split appears in Appendix E.

C Recognition feature contract and recurrent disclosure

At each station the four side temperatures (O, U, I, D) are mapped by Equation (1) of the main text to a_0, c_1, s_1, c_2. For each mode and station we compute the mean, standard deviation, final value, full-window slope, and recent-20 s slope, then append the six controls. The resulting 306-dimensional readout is target-free at inference. The matched raw-channel baseline uses the same five summaries on all 60 channels. The controls-only baseline uses only the six controls. Model probabilities are averaged within run before scoring, so each run contributes one paired decision and one ranking score. Linear and PCA baselines fit on train and validation members only, with PCA retaining 95% variance. Contract-matched controls-only ExtraTrees shares the candidate's 1000-tree, minimum-leaf, feature-subsampling, class-weight, and threshold contract. Gradient-boosting and random-feature baselines are specified in Appendix E.

Contrast-covariance extension. The descriptor uses the orthonormal zero-sum basis Q defined in the main text. Signed temporal summaries are computed from $Y = XQ^{\mathsf{T}}$ before temporal centering. Only the covariance uses $Z = HY$ and $C = Z^{\mathsf{T}}Z/(T-1) + 10^{-6}I_3$. The feature vector contains six controls, 225 signed-summary values, and 45 spectral values, for 276 features. With controls fixed, the whole vector is invariant to arbitrary additive stationwise common-mode drift. Only its spectral block is invariant to orthogonal contrast-basis changes. Side-dependent gains and missing sides fall outside these invariances. We use the same 1000-tree ExtraTrees settings, threshold, and three seeds as the primary readout. The artifact is `contrast_covariance_readout_20260922_v2.json`.

The GRU interface is a two-layer 128-unit recurrent model with a 32-dimensional latent and a GRUCell readout. Its prototype objective contains temperature, report-label, mode, and stress-proxy terms. The stress-proxy target is the post-window 290–310 s FE artifact injected by the window builder. It is used only as a training-time auxiliary target, never as an inference input or an event-time label. A target-derived future-temperature-difference regularizer is also training-only. The GRU report head is a multitask prototype trained with future-window auxiliary targets. Its sensitivity analysis is reported in this appendix.

Sampling and aliasing. The transform is exact for the four sampled temperatures. A continuous field $T(\theta) = \sum_m (a_m \cos m\theta + b_m \sin m\theta)$ aliases by residues modulo four. The mean contains $a_0 + a_4 + a_8 + \cdots$, while c_1 and s_1 mix higher odd harmonics. The $\sin 2\theta$ component is unobservable, and $\cos 4\theta$ is indistinguishable from a uniform offset. Thus c_2 is a fixed-phase second harmonic, not a complete continuous second-harmonic amplitude.

D Statistical protocol and inference audit

The primary endpoint is the geometry-harmonic candidate versus contract-matched controls-only ExtraTrees on the leave-pressure axis, using BA at the frozen 0.5 threshold. Every other model, split, and metric comparison is exploratory. The audit artifact `geometry_harmonic_statistical_audit.json` contains the paired rows. The main table also reports stratified run-bootstrap intervals for BA and AUROC.

Exact paired tests. On the leave-pressure axis the candidate records four wins and zero losses against both the raw-channel tree and the controls-only tree over 33 shared test runs, with 29 ties against raw. One seed gives three wins against controls and 30 ties. The exact McNemar probabilities are $p = 0.125$ against raw and $p = 0.125$, 0.125, and 0.25 against controls. These are tests of paired correctness, not direct tests of balanced-accuracy equality. The frozen-artifact BA difference is $+0.106$ against controls and $+0.143$ against raw. The run-bootstrap interval of the paired correctness difference is $[0.030, 0.242]$ against controls, but this interval is descriptive because the discordance count is small. The exact tests do not reach conventional significance. On leave-mass-flux all 15 runs tie against the controls-only tree, and 14 tie against raw.

Label-permutation nulls. Permuting the report labels across all 61 runs while holding split membership fixed (B=500 full refits of the three frozen pipelines per axis) places the observed candidate BA beyond the null on both axes. Leave-pressure gives 0.929 with $p = 0.002$, and leave-mass-flux gives 0.938 with $p = 0.002$. The candidate-versus-raw difference reaches $p = 0.090$ on pressure and is indistinguishable from zero on mass flux. Against contract-matched controls the differences are $+0.115$ ($p = 0.23$) and 0.000 ($p = 1.0$). These results show label information beyond the permutation null; the candidate-versus-raw contrast remains uncertain.

Threshold and calibration audit. Balanced-accuracy-versus-threshold sweeps (0.05–0.95) show the candidate at or above the raw-channel tree at every threshold in $[0.30, 0.60]$ on the leave-pressure axis (mean differences $+0.019$ to $+0.143$, with the curves crossing only at 0.65 at -0.017) and at every threshold in $[0.40, 0.70]$ on the mass-flux axis. Under the validation-threshold protocol (train-only fit, threshold chosen on the validation fold, test scored once) the ordering of Table 1 is preserved: leave-pressure 0.869 (harmonic) versus 0.702 (raw) and 0.610 (controls tree), and leave-mass-flux 0.917 versus 0.875 and 0.824. The gain therefore survives threshold selection, although the absolute operating points drop relative to the frozen train+validation fits. Calibration summaries are unstable at these test sizes. On the pressure axis the candidate has Brier 0.106 and 10-bin ECE 0.199 with mean probability 0.59 against an observed positive rate of 0.576, while the raw tree has Brier 0.134 and ECE 0.144–0.178 with mean probability 0.70. On the mass-flux axis the candidate has Brier 0.079–0.082 and ECE 0.134–0.138 versus raw 0.093–0.097 and 0.160–0.164. The candidate is better on Brier but *worse* on 10-bin ECE (0.199 versus 0.144–0.178 on the pressure axis) and better centered near the class base rate. Probability calibration is therefore incomplete, and thresholded statements use the fixed 0.5 contract.

Cluster bootstrap. A two-level cell-then-run bootstrap (10,000 resamples over the eight pressure–mass-flux cells) of the leave-one-cell-out audit gives a cell-cluster 95% interval of $[0.000, 0.071]$ for the candidate-over-raw difference (observed $+0.017$, one win and sixty ties over 61 unique runs): within the pooled envelope, the margin is small and its interval includes zero.

Block importance. Permuting each harmonic's 75 feature columns jointly on the held-out fold (200 repeats) attributes the largest mean balanced-accuracy drop on the leave-pressure axis to the circumferential mean a_0 (0.114), with the second harmonic c_2 second (0.088), c_1 and s_1 minor (0.051/0.048), and the controls negligible (0.007, reflecting the single training pressure level). On the mass-flux axis c_2 and a_0 are comparable (0.049/0.048). This post-hoc marginal view complements the refit-based component ablation (Appendix K): removing a_0 from the fitted full model hurts most, yet a refit with c_2 alone still recovers the full transfer balanced accuracy while a refit with a_0 alone does not (0.833). Both views agree that the operating controls contribute almost nothing on the extrapolation axis.

Label-flip sensitivity. Randomly flipping $k \in \{1, 2, 3\}$ report labels among all 61 runs and refitting the frozen contracts (300 repeats per cell) moves the leave-pressure headline gracefully: harmonic balanced accuracy 0.897 ± 0.035 / 0.869 ± 0.046 / 0.853 ± 0.049 for one/two/three flips (5–95% intervals down to $[0.761, 0.929]$), against raw $0.772/0.759/0.755$. The margin persists under realistic single-label noise, and harmonic remains above raw in every refit. Artifact: `geometry_harmonic_noise_audit.json`.

Controls-subset ablation. Controls-only trees using only (pressure, mass flux) reach 0.492/0.536 on the leave-pressure/leave-mass-flux axes. Adding heat flux and inlet quality raises them to 0.799/0.792. The full six controls reach 0.823/0.938. The two dimensionless numbers alone give 0.414/0.482. The mass-flux-axis tie is therefore driven by within-cell operating controls, especially heat flux and inlet quality, rather than by the two split axes themselves.

E CAPACITY-FROZEN EXTRA BASELINES AND SEED EXTENSION

All baselines share the frozen split/seed/threshold contract, and no hyperparameter is tuned. (i) Histogram gradient boosting: `HistGradientBoostingClassifier(max_iter=300, lr=0.05, max_depth=6, l2=1.0, balanced)` on harmonic and on raw summaries. (ii)

Random Fourier features: 300 cosine features of the standardized raw summaries with the median-heuristic bandwidth fit on training members only, plus controls, fed to the identical 1000-tree ExtraTrees. (iii) Random ReLU projection: 300 rectified features of the standardized raw summaries with fixed seeded weights, plus controls, same combiner. (iv) A 20-seed extension of the random split for the harmonic, raw, and controls-only trees. Leave-pressure balanced accuracies are 0.929, 0.786, 0.823, 0.857, 0.797, 0.750, and 0.774 for harmonic, raw, controls, GBM-raw, GBM-harmonic, RFF, and ReLU, respectively. Leave-mass-flux values are 0.938, 0.875, 0.938, 0.938, 0.812, 0.851, and 0.893 in the same order. Random split (20 seeds) gives harmonic 0.950 ± 0.068, raw 0.947 ± 0.051, and controls 0.590 ± 0.182—the random split is near-saturated for both temperature representations and separates them from controls only. Modern large sequence models (Transformer/Mamba/S4) are omitted because 61 runs with report-level weak labels do not support a fair capacity-tuning comparison. The GBM and random-feature rows provide contract-matched nonlinear comparisons at this sample size.

F PHYSICS-VALIDATION AUDIT

Full label-conditional statistics of the station-mean harmonics (medians, IQRs, two-sided Mann–Whitney probabilities, rank-biserial effects) are frozen in `geometry_harmonic_physics_audit.json`. Early (0–100 s) medians dryout/non-dryout are c_1 $-13.18/-6.84\,^\circ$C ($p = 0.0024$, $r = -0.45$), s_1 $+0.15/-0.04\,^\circ$C ($p = 0.0001$, $r = +0.58$), c_2 $+2.37/+1.01\,^\circ$C ($p = 0.0035$, $r = +0.44$), and a_0 $313.3/288.9\,^\circ$C ($p = 0.0002$). The late (290–310 s) field shows the same pattern (c_1 $-13.66/-6.87$, $p = 0.0017$ and c_2 $+2.39/+1.02$, $p = 0.0021$). The inner side is the hottest side in 30/31 dryout and 29/30 non-dryout runs in both windows, so the hottest side itself is not discriminative—the *magnitude* of the inner–outer asymmetry is. The late c_2 station mean correlates with $\log_{10}$ maximum von Mises stress at Spearman $\rho = 0.806$ ($p < 10^{-4}$, $n = 61$). These report-label associations support consistency with the proposed mechanism; causal identification requires independently reviewed physical-state labels.

G GEOMETRY-PRIOR IDENTIFIABILITY, SYNTHETIC WORLDS, AND CROSSOVER AUDIT

Identifiability. With sensors at $\theta_k = k\pi/2$, the projections a_0, c_1, s_1, c_2 are the exact discrete Fourier coordinates of the four *samples*. Under per-channel white noise σ^2 they have variances $\sigma^2/4, \sigma^2/2, \sigma^2/2, \sigma^2/4$. Monte-Carlo verification (20,000 trials, $\sigma = 0.3$) matches the theory to a maximum relative error of 1.4%. For the continuous field $T(\theta) = \sum_m (a_m \cos m\theta + b_m \sin m\theta)$ the aliasing algebra is explicit: mode m folds onto $m \bmod 4$, so $\mathbb{E}[\hat{a}_0] = a_0 + a_4 + a_8 + \cdots$, $\hat{c}_1$ mixes a_1 with $a_3, a_5, \ldots$, and $\hat{s}_1$ mixes b_1 with $-b_3, b_5, \ldots$ (verified numerically: $\cos 3\theta$ samples identically to $\cos\theta$, and $\sin 3\theta$ to $-\sin\theta$). The contrasts are unbiased for the continuous low-order coefficients only under band-limitation of those residue classes. Two structural facts follow: $\sin 2\theta$ is exactly zero at all four angles, so b_2 is unobservable by *any* four-sensor method and c_2 is a fixed-phase contrast rather than a full second-harmonic amplitude, and $\cos 4\theta$ evaluates to 1 at all four angles, so mode-4 content is observable only through the mean, where it is indistinguishable from a uniform offset.

Synthetic worlds. Four 56-run synthetic corpora mirror the real design (2 levels $\times$ 4 sub-conditions $\times$ 7 runs, labels balanced within each level, $+25\,^\circ$C level shift of the circumferential mean, per-channel noise $\sigma = 0.1\,^\circ$C, station amplitude profile). The frozen 1000-tree readout contract is applied unchanged. Leave-high-condition balanced accuracy (3 seeds): *mechanism world* (label doubles the inner-hotter c_1 to -13.2 and c_2 to $+2.4\,^\circ$C, condition-independent) harmonic 0.988, raw 0.500, controls 0.524. *condition world* (modes label-independent, label from a within-cell quality covariate) controls 0.940, harmonic 0.576, raw 0.500. *aliasing worlds* ($\sin 2\theta$ or $\cos 4\theta$ amplitude carries the label) all models 0.500–0.524. The two aliasing worlds must be read differently: a $\sin 2\theta$-driven label is invisible to any four-sensor method (a sensor-layout boundary), whereas a $\cos 4\theta$-driven label is observable through the circumferential mean and partially separable in pooled data (random-split balanced accuracy 0.60–0.64)—its failure under held-out conditions is the confounding of the mean channel with the $25\,^\circ$C level shift, the same mechanism that defeats raw channels in the mechanism world. The harness reproduces both real-data regimes—mechanism-driven transfer on the pressure

axis and condition-driven interpolation on the mass-flux axis—and separates true sampling blindness from condition-shift confounding.

Mode-level domain shift. Absolute standardized mean differences between the held-out level and the training levels of the real corpus (early context, station-mean modes): pressure axis a_0 1.70 (318.7 vs 279.9 °C), c_1 0.08, s_1 0.66, c_2 0.07, and mass-flux axis a_0 0.70, c_1 1.04, s_1 0.81, c_2 1.05. The pressure axis shifts the mean more than c_1 and c_2, while s_1 retains a moderate shift. Small marginal differences do not establish distributional invariance. The mass-flux axis shifts the contrasts themselves, consistent with the controls tie there.

Sample-efficiency crossover. Training budgets $n \in \{8, 12, 16, 20, 24, 28\}$ are drawn stratified 20 times each from the 28-run pressure-axis training pool. The test set stays the fixed 33 held-out runs. Balanced accuracy is harmonic ExtraTrees $0.793/0.831/0.855/0.868/0.908/0.920$, raw ExtraTrees $0.734/0.784/0.764/0.730/0.770/0.779$, raw MLP 0.726–0.748, and harmonic MLP 0.753–0.823. The harmonic prior has the highest mean at each sampled budget in this ranking diagnostic. The repeated subsamples are not independent validation, and no formal superiority claim is made.

H Energy-balance outlet-quality baseline

The classical zero-fit dryout rule for once-through steam generators evaluates the outlet thermodynamic quality from an energy balance over the heated length, $x_{\mathrm{out}} = x_{\mathrm{in}} + q''\pi d_{\mathrm{in}} L/(GA_{\mathrm{flow}} h_{\mathrm{fg}})$, with IAPWS-IF97 latent heats, $d_{\mathrm{in}} = 12\,\mathrm{mm}$ and $L = 4.936\,\mathrm{m}$ (the 14-turn centerline arc). Two threshold variants are reported. *Zero-parameter saturation rule* $x_{\mathrm{out}} \geq 1$: pooled balanced accuracy 0.887 (sensitivity 0.806, specificity 0.967), pressure-invariant across levels (0.875 at 3.5 MPa, 0.885 at 7 MPa). Leave-pressure 0.885, leave-mass-flux 0.652. *Train-fitted threshold* (selected on train+validation members only): pooled ranking AUROC 0.966, leave-pressure balanced accuracy 0.607 with AUROC 0.929, leave-mass-flux 0.938 with AUROC 0.875. The energy-balance physics carries most of the dryout ranking information, and its physical saturation threshold transfers across pressure by construction—the fitted threshold does not, because the operating point shifts with pressure. The harmonic readout's pressure-axis margin over the zero-parameter rule is $+0.044$. Appending x_{out} as a 307th feature to the harmonic contract leaves balanced accuracy unchanged on both held-out axes (0.929/0.938). The rule presumes the reported inlet quality and heat flux are calibrated channel values, which the data-readiness gates have not confirmed. Artifact: `outlet_quality_baseline.json` (`0343ca4cba21ccff7f092631f7e99cb9efad22c28965a4526b552b318d93cb20`).

I Attribution audit: common-mode separation, alignment, and fluctuation statistics

Because the four-mode projection is an invertible linear map, the transfer gain must be attributed rather than assumed. Four feature variants share the frozen tree/seed/threshold contract on the leave-pressure axis (harmonic 0.929, raw 0.786): **(i) Demeaned** channels (each side minus its station's circumferential mean per timestep, identical summaries, mean channel destroyed) reach 0.876—common-mode removal recovers most of the raw-to-harmonic gap. **(ii) Mean-preserving random rotation** (mean summaries kept as their own channel, and the three zero-mean contrast directions randomly recombined by 20 orthogonal draws) reaches 0.914 ± 0.031 with a 5–95% interval of $[0.863, 0.929]$: the empirical range overlaps the harmonic readout, with no formal equivalence or paired test in this 20-draw diagnostic. The earlier full-space orthogonal null (0.785 ± 0.053) mixes the mean into the contrasts. Once the mean is kept separate, *any* contrast basis performs comparably. The specific $c_1/s_1/c_2$ alignment therefore contributes at most a marginal increment, while the dominant factors are separation of the common mode and retention of contrast information. **(iii) Summary decomposition**: the five summaries restricted to the linear functionals (mean, last, slope, recent slope) give 0.875, while the standard deviation alone gives 0.929—the gain is carried by the *fluctuation statistics* of the contrasts, whose variance combines cross-channel covariances, not by the linear averages. **(iv) Classifier interaction**: gradient boosting on harmonic features (0.797) underperforms the same model on raw features (0.857), so the representation gain is readout-dependent and must not be described as representation-intrinsic.

Transfer-direction asymmetry and level-wise mass-flux holdouts. Reversing the pressure transfer (train on 7.0 MPa, test on 3.5 MPa) gives harmonic 0.917, raw 0.917, controls 0.694: the harmonic-over-raw advantage is *directional* ($3.5 \to 7$: 0.929 vs 0.786, whereas $7 \to 3.5$ is tied at 0.917). Holding out each mass-flux level in turn (harmonic/raw): 320 $0.714/0.857$, 530 $1.000/1.000$, 800 $1.000/0.976$, 1100 $0.917/0.875$—mixed ordering, with raw better at the lowest level. The mass-flux top-level tie (0.938) is therefore not an isolated anomaly but part of a mixed level-wise picture.

Condition-preserving label permutation. Permuting labels within each of the eight pressure–mass-flux cells (B=500 full refits) preserves cell-level label counts and condition strata while breaking run-level label–field association. The observed harmonic-minus-controls difference on the leave-pressure axis is $+0.115$ against a within-cell null of $+0.003 \pm 0.092$ (95th percentile 0.145): the observed value sits at the 88th percentile—positive in direction, not significant at conventional levels. This is the correct null for the incremental-information question, and it bounds the claim accordingly. Artifact: `geometry_harmonic_attribution_audit.json`.

J Classical-criterion null results

Fixed-threshold gates on Fr_c/Bo reach balanced accuracy 0.498 (bootstrap 95% CI 0.417–0.565). The univariate AUCs of $-\log_{10}\mathrm{Fr}_c$ and $\log_{10}\mathrm{Bo}$ against the report label are 0.485 and 0.498. A grouped cross-validated logistic diagnostic on $[-\log_{10}\mathrm{Fr}_c, \log_{10}\mathrm{Bo}]$ gives out-of-fold AUROC 0.408 (95% CI 0.267–0.553), balanced accuracy 0.426, and a label-permutation probability of 0.751. The expected coefficient signs persist in only 25% of training folds. The classical dimensionless criteria therefore do not separate this corpus at the report level, which is the motivation for reading the observed circumferential field. Artifacts: `criterion_audit.json`, `dimensionless_criterion_audit.json`.

K Grouping, transform-null, provenance, window, and cell robustness

Grouping audit. Replacing random partitions with five deterministic `StratifiedGroupKFold` folds over the 48 experiment-label groups, with mean test size 12.2, gives BA 0.957 ± 0.039 for harmonic, 0.916 ± 0.079 for raw, and 0.586 ± 0.139 for controls-only ExtraTrees. The corresponding AUROC values are 0.984 ± 0.037, 0.987 ± 0.019, and 0.686 ± 0.123. The group-held-out results are therefore a primary leakage sensitivity check. Thresholded recognition persists after group separation, while probability ranking is not uniformly improved.

Transform nulls. Replacing the fixed harmonic transform by a random orthogonal matrix or a fixed side permutation (controls, splits, trees, threshold unchanged) gives leave-pressure balanced accuracy 0.929 (harmonic) versus 0.786 (orthogonal) and 0.774 ± 0.021 (permutation). Leave-mass-flux 0.938 versus 0.875 and 0.875. AUROC remains mixed (pressure $0.917/0.902/0.898$, and mass flux $0.905/0.911/0.911$). A 100-draw random-orthogonal null places the observed pressure score at the 100th percentile ($p < 0.01$), but this full-space transform mixes the common mode into the contrasts. It therefore rejects the combined transform that fails to separate the mean, not a uniquely aligned geometry. The mean-preserving 20-draw rotation diagnostic provides no evidence of separation from the fixed basis.

Component ablation. With the 306-column budget held fixed, leave-pressure balanced accuracy is a_0 only 0.833, first harmonics only 0.850, c_2 only 0.929, a_0+first 0.857, full 0.929, raw 0.786. Leave-mass-flux is $0.917/0.938/0.938/0.938/0.938/0.875$. The transfer gain is carried by the second harmonic.

Label provenance. The primary report-index protocol contains 61 runs (31/30). A filename-candidate diagnostic changes two conflict groups (33/28). Excluding them leaves 59 runs (31/28). Under both diagnostic protocols the fixed classifier saturates at 1.000 on the held-out axes. Because the filename rule is not an independently reviewed label it remains a sensitivity diagnostic, and the direction of the result is driven by the two conflict groups. The two conflict run identifiers and all protocol counts are recorded in `geometry_harmonic_robustness_audit.json`. The two candidate errors on the leave-pressure axis are exactly these two conflict runs. Report-level labels record how the original experimental reports classified a run. Identifying them can at most *recover the historical reading rule* from the wall-temperature evidence. Whether that rule independently confirms the physical dryout state is a separate question that would require re-review of the original records and the evidence the reports relied on. The two are stated separately and not conflated anywhere in this paper.

Window sensitivity. Refitting on contexts 0–50, 0–100, and 100–200 s gives leave-pressure/leave-mass-flux balanced accuracy $0.911 \pm 0.015/0.938$, $0.929/0.938$, and $0.929/0.938$. AUROC is window-dependent ($0.926/0.911$, $0.917/0.905$, $0.919/0.887$). The thresholded pattern is stable across windows.

Leave-one-cell audit. Leaving out each of the eight pressure–mass-flux cells, fitting on the remaining cells, and scoring all 61 runs out of fold gives pooled balanced accuracy 0.934 ± 0.000 (harmonic) versus 0.918 ± 0.000 (raw) and AUROC 0.963 ± 0.002 versus 0.951 ± 0.001. The unique-run paired difference is $+0.016$ (interval $[0.000, 0.049]$, 60 ties, one win, no losses), with the cell-cluster interval of Appendix D. This bounds the pooled within-envelope margin and is the honest counterweight to the pressure-axis gain.

Cross-branch concordance. Pairing the three-seed leave-one-cell-out harmonic probabilities from the 0–100 s context with the $\log_{10}$ FE stress of the same run gives within-cell Spearman ρ of 0.738–0.952 (median 0.905) and a cell-mean-centered pooled $\rho = 0.882$ (run-bootstrap 95% CI $[0.785, 0.924]$, and cell-cluster CI $[0.828, 0.919]$). Both quantities share historical temperature/report provenance. The audit supports interface coherence, not causal or learned-cascade validity.

L CONDITIONAL FINITE-ELEMENT PROTOCOL, CONVERGENCE, AND CALIBRATION LEDGER

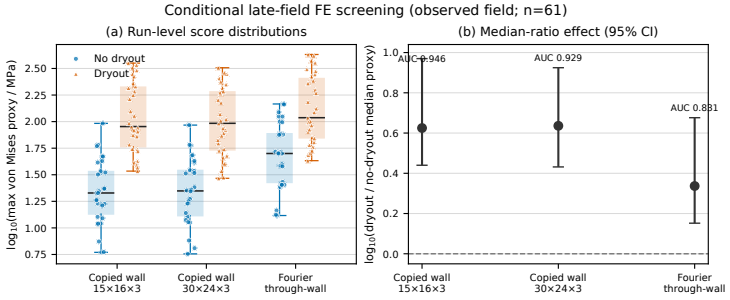


Figure 6: Conditional finite-element stress proxy by historical report label. The input field is the observed 290–310 s mean. This figure is an auxiliary numerical audit and not measured-stress validation.

The FE branch is an exploratory numerical proxy with no inference-time cascade. The reference implementation uses a Hex8 hexahedral mesh and receives the observed late field (290–310 s mean), geometry, and deterministic scenario boundary conditions, and solves Equation (3) of the main text. The GRU prototype does consume an FE stress artifact and a target-derived future-temperature term as training-only auxiliary supervision. This is a training dependency, not a deployment-time input or a claim that FE predictions drive the classifier.

Solver contract. The solver uses trilinear eight-node hexahedra with $2 \times 2 \times 2$ Gauss integration, isotropic Young's modulus $E = 195$ GPa, Poisson ratio $\nu = 0.30$, and thermal expansion $\alpha = 14 \times 10^{-6}\,\mathrm{K}^{-1}$. Nominal geometry uses coil diameter 112 mm, 14 turns, pitch 22.5 mm, and inner/outer radii 6/9 mm. The 5.82 m coordinate parameter maps station positions to the helix, whose geometric centerline length from the diameter, pitch, and turns is 4.936 m. These lengths are not interchangeable. The observed late outer-wall field is interpolated axially and periodically around the circumference and copied through the wall in the base scenario. The thermal reference is the interpolated surface-field mean. Six displacement constraints remove rigid-body modes, with a separately evaluated clamped inlet scenario. The score is the maximum von Mises value after volume-averaging Gauss-point stress tensors per element and averaging element values at nodes. It is not a Gauss-point peak stress. For the clamped scenario, referencing the field mean removes uniform absolute heating and makes the result conditional on that stress-free reference. The mesh has three radial nodes, with axial and circumferential resolutions specified per experiment. Sparse COO assembly and a direct sparse solve are used for refinement. These are conditional thermoelastic assumptions, not a calibrated two-phase-flow simulation.

Convergence disclosure. Against the nominal centerline geometry, the polyline arc-length error is 34.0%, 9.0%, 2.3%, and 0.56% at 30, 60, 120, and 240 axial nodes. In the new two-run refinement audit, the $30 \times 24 \times 3$, $60 \times 24 \times 3$, and $120 \times 32 \times 3$ per-run averaged-stress scores have median relative errors against the $240 \times 32 \times 3$ reference of 80.7%, 42.5%, and 25.7%, respectively. Extending the same two predetermined cases to $240/480/960 \times 32 \times 3$ reduces centerline arc errors to $0.561/0.140/0.035\%$. Relative to 960 nodes, the median stress-score errors are 33.17% at 240 and 8.81% at 480 (P95 $51.31/13.40\%$). These are axial

refinements with circumferential and radial resolutions held fixed, not a full three-dimensional convergence certificate. The two cases were selected deterministically, one from each report class. Geometry resolution improves while the averaged stress score remains mesh-sensitive. The historical 15-, 30-, and 60-node label rankings remain similar, with ROC-AUC $0.946/0.929/0.891$ and cross-mesh rank correlations 0.982–0.995. Deterministic geometry, through-wall, and support scenarios preserve rank correlations of 0.921–0.995, while absolute stresses vary substantially. All FE outputs are conditional and uncalibrated. They support relative screening under the stated Hex8, copied-through-wall, material, and support assumptions, not absolute stress, fatigue life, safety margins, or converged structural simulation.

Incremental value. Repeated five-fold stratified cross-validation (20 repeats) gives AUC 0.869 ± 0.107 for controls, 0.9656 ± 0.0675 for controls plus 13 late-temperature summaries, 0.9812 ± 0.0416 for controls plus FE stress, and 0.9667 ± 0.0670 for controls plus both. The repeat-mean FE-over-temperature increment is $+0.0011$ with a bootstrap interval $[0, 0.0025]$ that quantifies cross-validation Monte-Carlo noise, and the leave-one-condition mean increment is 0.000. The FE branch is therefore a conditional structural screen and coherence audit, not an independent predictive contribution.

Independent thermoelastic verification. A separate manufactured-solution harness assembles the Hex8 kernel and compares it with analytic isotropic elasticity without calling the production stress score. On three affine patch meshes, the maximum stress error is 6.56×10^{-7} Pa and the maximum relative elastic-energy error is 1.99×10^{-14}. Uniform heating on three curved helical meshes gives a maximum rigid-motion-fit displacement error of 2.67×10^{-12} m, with maximum residual von Mises stress 45.98 Pa and relative free residual below 3.7×10^{-17}. A linear temperature gradient with an analytic stress-free quadratic displacement field gives stress values 17.42, 8.71, and 5.81 MPa on three uniform meshes, with observed first-order convergence and maximum displacement errors below 3×10^{-19} m. This is numerical verification of constitutive assembly, thermal loading, gauge handling, and discretization. It does not validate measured stresses, material parameters, support conditions, or fatigue margins. Artifact: `thermoelastic_verification_20260922.json` (SHA256 `3838daad030021a4abb93238acb1f98300a2b44a38e79a14a0046bcb2`).

M DYNAMICS-BOUNDARY DETAILS

The matched recurrent probes are exploratory comparisons, separate from the primary pressure BA endpoint. Table 4 reports their complete pressure-held-out readout results. The logistic and ExtraTrees orderings differ, so these probes diagnose a representation–readout interaction rather than a uniformly superior latent state.

Table 4: Exploratory matched readouts on the pressure-held-out split. Entries are three-seed mean BA, with mean AUROC in parentheses. Learned representations use the final recurrent state.

Representation	Logistic ($C = 0.1$)	ExtraTrees (1000 trees)
Harmonic summaries	0.788 (0.876)	0.929 (0.917)
Raw-channel summaries	0.850 (0.902)	0.786 (0.901)
GRU latent	0.868 (0.941)	0.845 (0.939)
RSSM latent	0.500 (0.945)	0.500 (0.944)

Raw-unit magnitudes for the 100–200 s horizon: pooled $|\Delta T|$ median $0.041\,^{\circ}\mathrm{C}$ and P90 $0.529\,^{\circ}\mathrm{C}$ against the final context value. Per-run persistence MAE median $0.220\,^{\circ}\mathrm{C}$ and P90 $0.445\,^{\circ}\mathrm{C}$. 90.2% of runs change by less than $0.5\,^{\circ}\mathrm{C}$ on average. Direct temperature-head MAE over the same window is $0.606/1.739/0.703$ train-standardized units for random/pressure/mass-flux splits versus persistence $0.006/0.009/0.007$. The recurrent-input ablation (identical 2-layer/128-unit GRU contract, O/U/I/D steps replaced by a_0, c_1, s_1, c_2) gives run-level balanced accuracy $0.886 \pm 0.025/0.500/0.771 \pm 0.130$ (harmonic inputs) versus $0.943 \pm 0.052/0.500/0.813 \pm 0.108$ (raw inputs) and rollout MAE $0.634/1.765/0.731$ versus $0.620/1.765/0.712$. A 150-step fixed-state linear audit likewise does not beat persistence (history-only MAE $0.321/0.681/0.266\,^{\circ}\mathrm{C}$ versus persistence $0.210/0.237/0.208\,^{\circ}\mathrm{C}$). The horizon is nearly stationary. The fixed evaluation window provides limited evidence about strongly transient dynamics.

Learned-latent probe audit. We implement the partial-observation machinery as an RSSM with a deterministic 128-unit GRU core, a 32-dimensional stochastic state, an observation-conditioned posterior $q(z_t|h_t, o_t)$, and a prior $p(z_t|h_t)$. The diagonal-Gaussian KL is $\sum_j [\log(\sigma_{p,j}/\sigma_{q,j}) + (\sigma_{q,j}^2 + (\mu_{q,j} - \mu_{p,j})^2)/(2\sigma_{p,j}^2) - 1/2]$. We minimize the sum of one-step temperature MSE, report-label BCE, mode cross-entropy, and 0.1 times KL for 50 epochs with AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}). Input normalization uses train members only. Optimization forwards use train plus validation members only. Evaluation uses posterior means with no sampling, and each probe is fitted on train plus validation members and scored only on the held-out members. We compare a class-balanced logistic probe with $C = 0.1$ and a contract-matched ExtraTrees probe with 1000

trees, leaf size 2, square-root feature subsampling, and balanced class weights. In the corrected three-seed run, final-state logistic balanced accuracy on the pressure-held-out split is 0.500 for RSSM and 0.868 ± 0.040 for the matched GRU. Their AUROC values remain 0.945 ± 0.009 and 0.941 ± 0.009, respectively. The RSSM threshold failure therefore does not imply that its latent lacks ranking information. Held-out prior-mean rollouts have median MAE $32.88\,^{\circ}\mathrm{C}$ on the pressure split versus $0.233\,^{\circ}\mathrm{C}$ for persistence. The historical earlier RSSM artifact used an incorrect KL expression and summarized rollouts over both fit and held-out runs, so it is retained as a superseded development artifact and is not used for the corrected claims. The corrected artifact is `rssm_matched_readouts_20260922.json`, with the source hash and software environment recorded.

Physics-anchored residual RSSM audit. We also implemented a residual RSSM that keeps the last observed temperature field as an explicit anchor, predicts a low-rank modal increment, and differentiably decodes the predicted field through a fixed Hex8 temperature-to-stress operator. The operator contains 720 FE nodes, six stress components, and 60 temperature channels. We evaluated three seeds on random, pressure-held-out, mass-flux-held-out, and experiment-group-held-out splits for the full model, a no-FE ablation, and a no-anchor stability ablation. The full model has temperature MAE 0.214 ± 0.033 versus a persistence MAE of 0.210 on the random split, 0.240 ± 0.002 versus 0.237 on pressure transfer, 0.223 ± 0.014 versus 0.208 on mass-flux transfer, and 0.231 ± 0.020 versus 0.224 on group-held-out evaluation. The corresponding mean model-minus-persistence differences are positive on every split. Removing the anchor increases MAE to 24.0, 37.7, 31.8, and $30.3\,^{\circ}\mathrm{C}$, respectively. FE stress coupling lowers the stress error relative to the no-FE branch on the full matrix, but it does not produce a consistent temperature gain over persistence. We therefore retain this branch as an auditable physics-coupled extension and do not promote it to the primary method claim. The complete seed-level matrix and aggregation are in `results/physics_anchored_rssm_formal/summary.json` and `docs/reports/2026-09-25-physics-anchored-rssm.md`.

Table 5: Anchor ablation for the physics-anchored residual RSSM. Entries are temperature MAE in $^{\circ}\mathrm{C}$, reported as three-seed mean $\pm$ sample SD. Lower is better. The table is an exploratory historical ablation.

Split	Persistence	Full	No FE	No anchor
Random	0.210 ± 0.036	0.214 ± 0.033	0.210 ± 0.037	24.0 ± 5.9
Pressure-held-out	0.237 ± 0.000	0.240 ± 0.002	0.237 ± 0.000	37.7 ± 0.4
Mass-flux-held-out	0.208 ± 0.000	0.223 ± 0.014	0.211 ± 0.001	31.8 ± 0.3
Group-held-out	0.224 ± 0.023	0.231 ± 0.020	0.225 ± 0.023	30.3 ± 5.8

The no-anchor column diagnoses rollout stability rather than forecast utility. Explicit observation anchoring prevents catastrophic absolute-temperature drift across every split family, while the full model remains slightly worse than persistence in temperature MAE.

N Compact metric ledger

Entries are random / leave-pressure / leave-mass-flux unless noted. Zero standard deviations on the held-out axes reflect the deterministic test set, not evidence of stability.

Table 6: Metric ledger. GBM = histogram gradient boosting. RFF/ReLU = random-feature maps of the raw summaries with the matched ExtraTrees combiner.

Metric	Random	Pressure	Mass flux
Harmonic ExtraTrees BA	0.967 ± 0.058	0.929	0.938
Harmonic ExtraTrees BA (20 seeds)	0.950 ± 0.068	–	–
Raw-channel ExtraTrees BA	0.967 ± 0.058	0.786	0.875
Raw-channel ExtraTrees BA (20 seeds)	0.947 ± 0.051	–	–
Controls-only ExtraTrees BA (matched)	0.616 ± 0.077	0.823 ± 0.015	0.938
Controls-only logistic BA	0.806 ± 0.191	0.893	0.813
GBM raw BA	0.910 ± 0.016	0.857	0.938
GBM harmonic BA	0.910 ± 0.016	0.797	0.812
RFF ExtraTrees BA	0.933 ± 0.058	0.750 ± 0.036	0.851 ± 0.041
ReLU ExtraTrees BA	0.933 ± 0.058	0.774 ± 0.021	0.893 ± 0.039
Harmonic candidate AUROC	0.960 ± 0.069	0.917	0.905
Raw-channel AUROC	1.000	0.901	0.917
GRU report BA (prototype)	0.943 ± 0.052	0.500	0.792 ± 0.072
GRU circumferential-mode macro-F1	0.575 ± 0.020	0.192 ± 0.006	0.820 ± 0.040
Free-rollout MAE / persistence	$0.606/0.006$	$1.739/0.009$	$0.703/0.007$
FE ROC-AUC by mesh (15/30/60 axial nodes)	0.946 / 0.929 / 0.891		

O ARTIFACT IDENTITY LEDGER

The following SHA256 values identify the internal artifacts bounding the historical claims. Source workbooks are not redistributed.

Artifact	SHA256
geometry_harmonic_classifier.json	543cd33983b47111b5f3469f028c7dce9afd509ee6b8d9f1693d41b385fe6249
geometry_harmonic_statistical_audit.json	b4085bc9a05df7ae19d92480094f429e5a8849741e6e09e4c525df494a846fdb
geometry_harmonic_physics_audit.json	fc5df7527f2c4995916d3656d44ec5c49cf31b29ffcef77c913df21580969d00
geometry_harmonic_extra_baselines.json	8a2a31f570f6491cb19f5c70d553368fd186a25d5a19f336df86c1cfb40bbabf
conditions_and_rollout_context.json	bb5daf3519d882bf1a1e148b2ea1fe642af04960b7aa4e267403bc67dcdabf34
geometry_prior_audit.json	38f212a5af50221c501ef3f05a11b17dd3d28112ec80e1623f627c32f1ea6536
geometry_harmonic_noise_audit.json	a56e1b0cdfc1127b5e4ae03c8833f027ed22f8d56d12ce5b95a37b8b8488d49b
geometry_harmonic_attribution_audit.json	768f4fa11226fef421df11c86272d8696b3199f9d36edf1e0e928cfa53fa169d
outlet_quality_baseline.json	0343ca4cba21ccff7f092631f7e99cb9efad22c28965a4526b552b318d93cb20
geometry_harmonic_baseline_audit.json	a371ea51907c1d0c2e5040ad0146d1e6618d885f4ab162cefddea629c2620ee8
geometry_harmonic_linear_baselines.json	397bd323e082a83c0da6d64b636ecc3359ff13be9fa214c0ace8bb40265c5466
world_model_classification_audit_highres.json	4cfcf1dd63441d4f91fbb1acc7a3fc8dde91a01efda13981aaf09a06968c4564
risk_separation_audit.json	4549748743edd410b20920a4d76e545004f39af8c4a8008f2378225266876eb4
geometry_harmonic_cell_audit.json	2f641d6b07a70628ca47031c623ff29b46674fe2f55445f4a0de9d854ddc5dbc
geometry_harmonic_window_sensitivity.json	ea3e2d14022730a1e124666a26374af9cc5785771af847610abe474102f3ec94
cross_branch_concordance_audit.json	c27fd80f5b9b194465ead9a2940fe77f992f8d10a646329db5d4ea09bc26c783
geometry_harmonic_capacity_audit.json	2d28588a040d581f01f9c3918e2bd012878c960d456cac83e0d0bae9bcaed629
geometry_harmonic_robustness_audit.json	dc820904f4ed7bbdc2ed21d6cd6b496604f554753e0dc9749e1d1b84a1c66a90
rssm_matched_readouts_20260922.json	01fdba1e38bebb3937852d712947b8427e16f1c413cd804c9912a7aad21dd5d5
condition_adjusted_physics_20260922.json	63c8bd91ac3a4574faa1f666ca868eb9b06eb609009c5dcc25585e3e2b2db660
dispersion_baselines_20260922.json	a925e951c9beaf7b09afac79d033504c1ef2f7917097a736317da53e40e12e6b
fe_2run_960.json	7ca2326e886b8cea09a5511bb5a3048e27dfcdd3ee9e38875c766a155a5a67db
contrast_covariance_readout_20260922_v2.json	02ef017a1d64476a51a26a0e22bdff430ce89cdfb764ad34e4c88e9f4e81434a
thermoelastic_verification_20260922.json	3838daad030021a4abb93238acb1f98300a2b44a38e79a14a0046bcb2

P GENERATED-PROBE REHEARSAL

To exercise the prospective-evaluation machinery before any real holdout exists, three program-generated probes (deterministic generator, seed 20260918, with a format identical to the DAQ export) were scored with the frozen 2026-09-15 classification bundle and the candidate multi-sensor event detector. The probes are the kit's reference cases: a dryout run at a *third* pressure level (5.25 MPa, the midpoint of the two historical levels, with onset at 230 s and inner-side first), a same-condition non-dryout control at 0.10 lower inlet quality (censored at 469 s), and a new-geometry apparatus variant (13 mm bore, 177 mm coil) with an upper-side-first onset at 250 s. These generated probes exercise file parsing, frozen inference, and event-detection interfaces with explicit generated provenance.

Outcome. (i) *Boundary pair*: none of the three frozen classifiers flags the non-dryout probe (harmonic/raw/controls probabilities $0.197/0.042/0.373$)—no false alarm at the new pressure level. (ii) *Pre-event horizon*: the dryout probes carry onsets 130–150 s after the end of the frozen 100 s context window, and no classifier exceeds 0.24 probability there—consistent with the paper's explicit no-early-warning boundary. The probes quantify how little pre-event signal the contract's context carries. (iii) *Event detection*: the multi-sensor MAD detector at 10σ (20 s sustain) detects the two dryout probes 7 and 9 s after ground-truth onset with *correct first channels* (TW13-I and, on the variant apparatus, TW12-U—the upper-side-first mode absent from the historical corpus). At 15σ the non-dryout probe is false-alarm-free at the cost of 11–13 s lag, whereas thresholds of 5–10σ produce false alarms. The lag/false-alarm trade-off and the circumferential first-direction machinery are the transferable diagnostics. Synthetic-probe classification scores are reported only as diagnostics and are not compared with historical metrics. Artifact: `synthetic_probe_rehearsal.json` (`680eb3f46cb5ad00eaa6df7e052d0ad886ae820221545465484de213a8a87fc4`).

Batch A: four boundary pairs at 5.25 MPa. A generated batch (deterministic seed 20260919, with critical quality interpolated from the 21 real boundary pairs, dual-review event tables, and all files SIM-prefixed and `simulated`) supplies four dryout/non-dryout boundary pairs spanning the four mass-flux rows (320–1100 $\mathrm{kg\,m^{-2}s^{-1}}$), with onsets 190–230 s and a mode mix of centrifugal (inner-first $\times 2$), upper-side-first, and a lower-side-first film-reversal mode absent from the historical corpus. Frozen-bundle outcome: **zero false alarms** on all four non-dryout probes for every model. The controls rule orders all four pairs correctly (0.51–0.56 versus 0.42–0.47) while the wall-field readouts stay flat (≈ 0.20–0.23)—onsets lie 90–130 s beyond the frozen 100 s context, reproducing the pre-event horizon and the condition-driven/controls-interpolation regime of the synthetic-world audit. The multi-sensor detector at 10σ finds all four onsets with mean lag $+6.0$ s (range $+4$ to $+10$ s) and **correct first channel in 4/4 cases** (TW14-D, TW12-I, TW11-I, TW13-U), including the two first-dryout modes the historical corpus does not contain. At 15σ all non-dryout probes are false-alarm-free (lags $+6$ to $+13$ s), whereas 10σ produces false alarms against a 0.3–0.4 K synthetic noise floor, quantifying that trade-off. These numbers diagnose the frozen machinery on synthetic probes at an unseen pressure level and are reported separately from historical accuracies. Artifact: `synthetic_probe_rehearsal_batchA.json` (`4353bdfa4610657f0d71774d8471c0c8748b4b220a581a0748b5eb2a8ea32d34`).

Q Prospective confirmation protocol

Before accessing prospective outcomes, freeze models, input processing, metrics, threshold, and baseline comparisons with hashes. New classification runs must have disjoint acquisition identities and independently reviewed report labels. Event detection additionally requires reviewed onset or censor intervals, and engineering stress requires calibrated thermal and mechanical observations. Our existing integrated runner requires the full event, calibration, energy, and FE contracts and remains closed. Its eight-run intake minimum is an engineering pilot size, not a power calculation or sufficient evidence of superiority. Before collecting a confirmatory sample, specify the target effect and choose its size by an appropriate power or precision analysis. Evaluate each released sample once and report paired uncertainty regardless of outcome. Raw workbook release remains subject to data-owner authorization.

R Development generated prospective-shaped experiment

To exercise the complete prospective intake and engineering-data interfaces, we ran a separate 24-run generated package at `data/supplementary_synthetic_20260925`. The package contains 12 event traces and 12 censored traces across three operating groups (4.25, 5.25, and 6.25 MPa), with controlled clock perturbations of -150, 0, $+100$, and $+250$ ppm. All 24 XLSX payloads have distinct content hashes. We process this package as a development experiment using the same file, clock, calibration, and stress-chain interfaces as the planned release. It tests the prospective interfaces without providing a real prospective holdout. The package-level evidence boundary and generated-reference limitations are stated together in the final limitations section.

Calibration-chain diagnostic. The package supplies 44,640 repeated generated calibration observations for 1,488 run–channel pairs, covering 60 wall-temperature channels plus pressure and flow. Fitting an affine response with a post-stage drift indicator gives a median residual RMSE of 0.0282 and a P95 of 0.0346 in native channel units. Recovery errors relative to the generator parameters are 6.08×10^{-5} in gain, 0.00604 in offset, and 0.00860 in end drift. These values verify the parser, units, affine inversion, and stage bookkeeping on generated references. The reference uncertainty is declared zero by construction, and there is no certificate, instrument serial number, or laboratory calibration record. Measured calibration requires instrument records and traceable reference data.

Thermoelastic stress-chain diagnostic. The package supplies 34,560 generated strain/stress rows from three stations in each run. Reconstructing stress from the generated indicated strain gives bias -0.00627 MPa, RMSE 0.3896 MPa, and P95 absolute error 0.7391 MPa. The reference stress and reconstructed stress share the same idealized uniaxial law with $E = 180$ GPa, $\alpha = 16 \times 10^{-6}\,\mathrm{K^{-1}}$, and a prescribed restraint fraction. This constitutive round-trip test covers the idealized uniaxial law used by the generator; it does not include pressure loading, coil bending, creep, plasticity, support conditions, or through-wall thermal gradients.

Generated interface check. The package passes checks for monotone time, finite values, identity, clock, calibration format, and constitutive consistency. Its 24 unique hashes prevent accidental duplicate payloads within the package. The complete machine-readable result is `results/metrics/supplementary_synthetic_analysis_20260925.json`, with the compact table in `paper/tables/table_supplementary_synthetic_audit.md`.

S CONDITION-ADJUSTED CIRCUMFERENTIAL ASSOCIATION

We average each mode over the 0–100 s window and 15 axial stations, then compute the dryout-minus-nondryout mean difference within each pressure–mass-flux cell. The primary statistic gives each of the eight cells equal weight. Across a_0, c_1, s_1, c_2, effects are $+23.989, -8.001, +0.264, +1.652\,^\circ$C. Positive/negative direction counts are $8/0, 0/8, 7/1, 6/2$. We use 10,000 shared within-cell label permutations, retaining each cell's class counts, and the two-sided Monte Carlo test $(1 + \#\{|T_b| \geq |T_{\text{obs}}|\})/(B + 1)$. Each raw p is $1/10001$, with Holm-adjusted $p = 4/10001$ across four modes. These associations condition on pressure and mass flux, not on heat flux or inlet quality. The shared provenance limits this analysis to association with report labels. The artifact is `condition_adjusted_physics_20260922.json`.

Scope of the historical comparisons. The pressure experiment evaluates one $3.5 \to 7$ MPa direction on one apparatus. Exact paired tests do not establish a significant incremental classification gain. Within-cell permutations put the harmonic-over-controls increment at the 88th percentile. The leave-one-cell improvement is one corrected run among 61, and the two pressure errors coincide with conflicting label provenance. These retrospective analyses reuse the existing corpus; independent confirmation requires new acquisitions. Sensor calibration and energy closure are unavailable.

Generic dispersion comparator. For each time and axial station, replace the four sides with their mean, population standard deviation, range, and mean absolute deviation from the mean. Apply the same five temporal summaries and append the same six controls, giving 306 features. This transform is invariant to side permutation and discards orientation. The raw, harmonic, and dispersion pipelines share seeds, split membership, 1000-tree ExtraTrees settings, and threshold 0.5. Generic dispersion achieves balanced accuracy $0.967/0.860/0.938$ on random/pressure/mass-flux splits, with pressure SD 0.031. Its pressure AUROC is 0.920, versus 0.917 for harmonics. The feature count and classifier settings are matched, but effective feature dependence differs. Artifact: `dispersion_baselines_20260922.json`.

Contrast-covariance results. Balanced accuracy/AUROC are $0.967/0.960$, $0.876/0.896$, and $0.938/0.881$ on random, pressure-held-out, and mass-flux-held-out splits. Pooled leave-one-cell-out scores are $0.951/0.958$, compared with $0.934/0.963$ for harmonic and $0.918/0.951$ for raw summaries. At each seed, covariance wins on one run and loses on none against harmonic ($p = 1.0$), and wins on two and loses on none against raw ($p = 0.5$), using exact two-sided McNemar tests. Seeds reuse experimental units and are not independent replications. These comparisons support a small descriptive cell-held-out gain, not statistical superiority or isolated attribution to covariance spectra. The descriptor also changes mean retention and feature dimension.

T CONTRAST-COVARIANCE INVARIANCE AND SENSOR STRESS

For $X \in \mathbb{R}^{T\times 4}$, let $Y = XQ^\mathsf{T}$ and $H = I_T - \mathbf{1}_T\mathbf{1}_T^\mathsf{T}/T$. For any $b \in \mathbb{R}^T$, $(X + b\mathbf{1}_4^\mathsf{T})Q^\mathsf{T} = Y$ because $Q\mathbf{1}_4 = 0$. Every deterministic summary of Y and its regularized covariance is therefore unchanged. Appended controls must remain fixed. A frozen deterministic classifier then receives the same feature vector, up to floating-point arithmetic.

For an orthogonal $R \in \mathbb{R}^{3\times 3}$, replacing Q with RQ gives $Y' = YR^\mathsf{T}$ and $C' = RCR^\mathsf{T}$. Trace, trace of the square $\operatorname{tr}(C^2)$, and determinant are unchanged under this similarity transform. Coordinatewise means, final values, slopes, and standard deviations need not be unchanged. This proves invariance of the spectral block, not of the full feature vector, to contrast-basis rotations. A side-dependent gain generally maps common-mode temperature into the contrast subspace, and imputation alters the observed contrasts. Neither operation is covered by the offset result.

Physical tradeoff. Common-mode rejection can discard a genuine target signal if dryout raises all four sensors equally. Appending controls can also reintroduce condition dependence. Exact representation invariance therefore does not imply distributional invariance, label preservation, or calibration. The primary harmonic descriptor retains the mean and does not share the full descriptor's offset guarantee.

T.1 SPLIT-SPECIFIC SENSOR PERTURBATIONS

The model is fitted on clean train+validation members and evaluated on perturbed test inputs without refitting. Controls are unchanged. The common-mode test adds the same $2\,^{\circ}\mathrm{C}$ sinusoidal drift to all four sides at each station. The gain test applies $T_s' = 300 + g_s(T_s - 300)$ in degrees Celsius with $g = (1.05, 0.95, 1.02, 0.98)$. The missing-side test drops the bottom (D) side at every station and fills it with the simultaneous mean of the remaining three. These diagnostics model declared perturbations rather than experimentally estimated sensor-error distributions.

Table 7: Contrast-covariance test-input perturbations. Entries are three-seed mean balanced accuracy, followed by mean absolute probability change relative to the same fitted model's clean predictions. Splits have different held-out populations and are reported separately.

Perturbation	Random ($n = 10$)	Pressure ($n = 33$)	Mass flux ($n = 15$)
Clean	0.967 (0.000)	0.876 (0.000)	0.938 (0.000)
Common-mode drift	0.967 (0.000)	0.876 (0.000)	0.938 (0.000)
Relative side gain	0.967 (0.050)	0.911 (0.023)	0.938 (0.048)
Bottom-side dropout	0.967 (0.100)	0.753 (0.129)	0.938 (0.079)

The unweighted across-split means of 0.938 for gain and 0.886 for dropout conceal the pressure-specific behavior. The means are descriptive and are not pooled-run performance. The gain perturbation improves this thresholded pressure score, but it should not be interpreted as a calibration correction. This stress protocol evaluates contrast covariance only, so it does not compare sensor robustness across representations.

U RETROSPECTIVE ACQUISITION AND EVENT-REVIEW READINESS

A fresh source-workbook audit recovers 74 parseable files representing the same 61 unique content hashes. All parseable time axes are monotone. Among 3,660 run–wall-channel pairs, no missing values or exactly constant finite channels were observed, and the largest sampling interval was 2.495 s. These checks establish acquisition usability; sensor accuracy and channel calibration require dedicated measurements.

Anonymous trace exports and two blank review sheets were prepared for all 61 runs. No completed independent event reviews or adjudication were available. A baseline-relative detector flagged 19 runs at threshold 10 with a 20 s window, including 14 of 31 report-positive and 5 of 30 report-negative runs. These are label-stratified candidates. Event sensitivity and false-alarm rates require independent event reviews; runs that begin during dryout may be left-censored.

V LIMITATIONS AND EVIDENCE BOUNDARY

The generated package begins at an arbitrary 2000-01-01 simulation epoch, derives its thermal trajectories from historical exemplar means rather than a two-phase solver, and uses scripted event times and right-censoring. Its calibration references are exact simulator inputs with zero declared reference uncertainty. Its stress target is an idealized local thermoelastic coupon calculation rather than an independent mechanical observation. This package is a development test of parsing, clock handling, calibration bookkeeping, and constitutive round trips. Real prospective generalization, measured sensor calibration, and measured-stress validation require disjoint acquisitions, independently reviewed event and censor intervals, traceable instrument calibration, energy closure, and calibrated mechanical and thermal boundary conditions.