# FROM SCORES TO ACTIONS: CERTIFIED PROMOTION AND $K > 2$ CONTROL OF JOINT TIME-SERIES GENERATORS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Joint time-series generators should create new trajectories while allowing relationships to evolve and protecting what a task requires. We call this goal *selective structural fidelity*. It raises three distinct questions: are the generated paths plausible, is a requested change executed correctly, and is there evidence to adopt it? We develop a reference-aware evaluation contract that separates these questions. Our rank-adaptive $K > 2$ path controller (KRPC) realizes matrix requests through thin products, with endpoint and complement preservation, basis covariance and explicit refusal. The evidence distinguishes execution from benefit. On a locked 2024 weather panel, event-preserving correction improves six quality measures over its calibrated reference; path conditioning also improves Energy and increment Energy over a matched history-conditioned control. A distinct, locked 2025 expert policy improves Energy by 0.0600% over its protected reference, but does not lead the full roster. Rank-4 requests on $K = 500$ streamflow ensembles establish high-dimensional execution. EEG and controlled decision studies expose quality–coverage and information limits. These are task-specific gains, not universal dominance.

## 1 INTRODUCTION

A joint time-series generator should create new trajectories, not copies. Channels may become more volatile together, separate, and later recouple. A user may want to change their common motion while retaining a spread, an accumulated total or a specified event. The task must distinguish *what may change from what must remain*.

We call this *selective structural fidelity*. Its object is an ensemble of joint paths, not a point forecast. A request names an editable relation; a protection statement names a property and its reference. That reference can vary with context: preserving structure during an action does not impose stationarity across time. Figure 1 shows changed path shapes with unchanged contrasts and endpoints.

Three questions follow: is the base law plausible, is the action correct, and is using it justified? One score cannot answer all three. A deliberate change to a correct law need not improve Energy against unrequested observations (Gneiting & Raftery, 2007); a better average score need not preserve the relations needed for the task.

Existing models address general time-series generation (Yuan & Qiao, 2024), conditional imputation (Tashiro et al., 2021), and multivariate future-path generation (Rasul et al., 2021; Shen & Kwok, 2023; Drouin et al., 2022; Kan et al., 2022). Generative foundation models such as Sundial also sample future trajectories, although its univariate pretraining does not explicitly model cross-variable dependence (Liu et al., 2025). Constrained generation is also established: COP post-processes outputs and GuidedDiffTime guides sampling (Coletta et al., 2023). Learn then Test calibrates multiple risks under explicit assumptions (Angelopoulos et al., 2022); transformed proper scores target relevant features (Pic et al., 2025). Regions of Reliability already studies the finite-sample discrimination of multivariate scores (Marcotte et al., 2023). Our contribution is not another universal score or the first constrained sampler. It is a reference-aware task specification, a concrete high-dimensional realization, and evidence that separates their different responsibilities.



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.

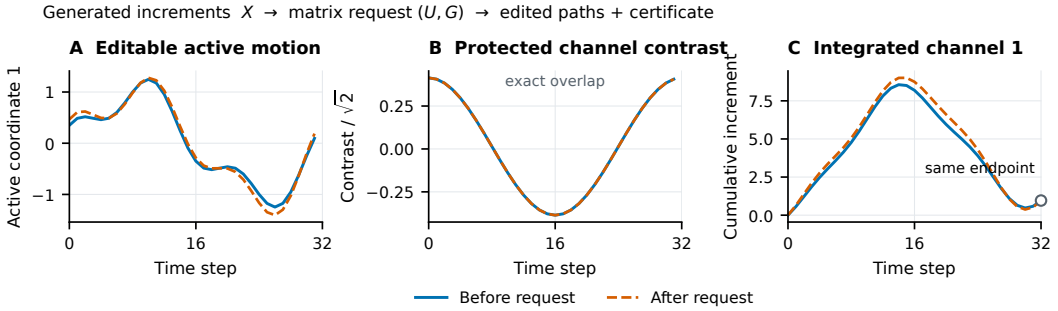


Figure 1: **Change the shape; preserve the specified structure.** Deterministic illustration of Equation 3, not an experimental sample. Here $K = 3, H = 32$ and $G = R_{\pi/6}\,\mathrm{diag}(1.15, 0.85)R_{\pi/6}^{\top}$, where $R_\theta$ rotates the active plane. The request changes active coordinates. The protected coordinate is unchanged at every step; the integrated path returns to the same endpoint. The certificate establishes these identities, not similarity to the true law or downstream benefit. A production request additionally needs frame, support and action checks.

**Two contributions.** Our evaluation contribution is a task specification that separates distributional quality, request execution and justified adoption. It names protected references, tests whether measurements detect declared failures, and distinguishes primary benefit from permitted secondary costs. Our method contribution is KRPC: a thin matrix-request compiler with exact structural identities, basis covariance and explicit refusal. Composition and observability results explain limits of these guarantees. We organize the evidence around four questions: protected quality on weather paths, quality–coverage trade-offs on EEG, high-rank request execution on streamflow, and the value of relational information for a controlled decision. These studies use different operators under one task specification; they are not demonstrations of one model dominating every domain.

## 2 What must a controlled generator establish?

For a reference path or ensemble $x$, let $\Phi(x)$ be its protected properties and $\Psi(x)$ its editable coordinates. An action remains in $\mathcal{F}_\Phi(x) = \{z : \Phi(z) = \Phi(x)\}$ while realizing the request within declared tolerances. The task names a primary loss and admissible secondary costs. A weather correction, for example, may edit path shape while retaining memberwise exceedance indicators and the ensemble mean. These differ from the endpoint/complement protections in Figure 1; neither set asserts that its reference is the unknown true law.

Let $Y \in \mathbb{R}^{H\times K}$ denote a joint path with $H$ time steps and $K$ synchronized channels. A generator produces $M$ paths, optionally conditioned on information $c$. In our conditional-generation experiments, $c$ is an observed history and $Y$ is the subsequent observed path. We first declare whether the scored object is a latent conditional law or the finite ensemble actually issued. For conditionally iid draws $X_{1:M} \sim Q_\theta(\cdot \mid c)$, independent of the verifying future given $c$, the fair Energy estimator is (Gneiting & Raftery, 2007; Ferro, 2014)

$$\widehat{\mathrm{ES}}(X_{1:M}, Y) = \frac{1}{M}\sum_m \|X_m - Y\|_2 - \frac{1}{2M(M-1)}\sum_{m\neq m'} \|X_m - X_{m'}\|_2. \tag{1}$$

The first term measures discrepancy from the observed path; the second accounts for ensemble dispersion. Neither term alone establishes calibration. Our newer coupled-ensemble studies instead score the issued measure $\widehat{Q} = M^{-1}\sum_m \delta_{X_m}$, replacing the second denominator by $2M^2$. This is its exact Energy score, not an unbiased estimate of a member's marginal-law risk. Historical off-diagonal scores retain their original benchmark interpretation when iid sampling is not established.

Uncertainty must also respect the unit of evidence. An *outer unit* is the group reduced to one paired comparison before inference. For EEG, it is one subject after averaging eight origins, not one member, channel or origin. Chronological units may still be dependent. Appendix B gives the coupling-bias identities and the assumptions behind each conditional interval.

The following standard consequence of proper scoring (Gneiting & Raftery, 2007) motivates the task; it is not a new scoring theorem.

**Theorem 1** (Standard Energy-risk identity: quality–control boundary)**.** *Let $P$ be the true zero-request conditional path law and let $Q = T_{\#}P$ be the law after a nonidentity request, both with finite first moments. If their Energy distance $D_{\mathrm{E}}(P, Q)$ is positive, then*

$$\mathbb{E}_{Y\sim P}[S_{\mathrm{E}}(Q, Y) - S_{\mathrm{E}}(P, Y)] = \tfrac{1}{2}D_{\mathrm{E}}(P, Q) > 0. \tag{2}$$

*Thus a valid control that changes a correct generative law cannot be expected to improve Energy against observations from the unrequested law. Assessing distributional fidelity after a request requires a justified request-matched evaluation law, as in a known-truth intervention. Fixing a request before the outcome is observed prevents leakage but does not supply that law.*

This standard identity fixes the division of labor. We score the base law at the identity request. For nonidentity requests we test tracking, protection and refusal, adding distributional checks when a justified request-matched law exists. A covariance constraint alone defines a feasible class, not a unique true counterfactual law. Appendix H gives the proof. Correcting a misspecified base remains legitimate, but its quality gain is separate from its certificate (Section 4.1).

We use *promotion* to mean accepting a candidate under a declared contract. Each mandatory check must pass; a gain elsewhere cannot cancel a failure. Before evaluating a held-out panel, the contract fixes sources, eligible cases, comparison units, primary benefit, protected properties and allowed costs. Exact algebraic protection and statistical evidence are different assurances. Moreover, not every quality or non-copying endpoint has an independent detection-power study. *Certified* refers only to the stated, tested contract, not universal safety (Appendix D).

**Primary success need not mean winning every score.** The primary objective determines benefit; mandatory protections determine admissibility. Secondary outcomes may worsen within declared tolerances and must be reported. Thus a tail-focused generator can trade some average Energy for a better tail score. A lower primary loss cannot compensate for breaking a required invariant, and a descriptive trade-off does not change an earlier protocol's decision.

**Qualify the measurement before ranking generators.** A population-identifying score can have low power at the available ensemble size. Conversely, a detector can reject harmless new paths too often. We therefore test a measurement against both named failures and legitimate variation, fixing its threshold before the challenge draws. A stage-dependent relation is a useful example: positive early correlation and negative late correlation can cancel in a pooled covariance. Flattening both stages then looks faithful to the pooled summary while erasing the intended dynamics. The task must supply the relevant stages and channels to every comparator, rather than reward a feature available only to our evaluator.

Our independent-draw comparison uses 3,072 calibration and 1,536 challenge trials from three declared processes, with eight statistics sharing the same ensembles. Phase–lag moments achieve 80.396% average dynamic detection, with a maximum simultaneous false-positive upper bound of 3.423%. They establish the prespecified advantage over four references, but not over phase-pooled variogram or Energy on the identical dynamic features. A common mean shift supplies the complementary failure: full-path Energy detects it, whereas the centered relation statistic is largely insensitive. These findings support task-matched qualification, not a universally superior score (Appendix Z).

This evaluation is distinct from model selection. New challenge draws test the measurement on the declared process families; they do not make an already-used real panel fresh. A deployment claim also needs the complete eligible-case denominator, including refused requests and their fallback outcomes. Otherwise, a rule that declines every action can appear safe without delivering useful generation.

## 3 Generating paths and controlling selected structure

KRPC edits generated paths in active channel directions while retaining their complement. Time centering makes the edit sum to zero, retaining the integrated endpoint. The request specifies an active response, not a trajectory.

**A worked request, not a new forecasting target.** Take $K = 3$, $r = 2$, $u_1 = (1, 1, 0)^\top/\sqrt{2}$ and $u_2 = (0, 0, 1)^\top$. A non-scalar SPD request changes the common motion of channels 1–2 together with channel 3. It retains the contrast $(x_1 - x_2)/\sqrt{2}$ at every step and each cumulative endpoint. This specifies how to act, not why a user should prefer that covariance.

Let $X \in \mathbb{R}^{N\times H\times K}$ be generated increments, $P_H = I_H - H^{-1}\mathbf{1}\mathbf{1}^\top$ the time-centering operator, and $U \in \mathbb{R}^{K\times r}$ a qualified orthonormal frame with $r \ll K$. Here $N$ is the number of acted-on paths ($N = M$ for one ensemble). The reported policy chooses the largest $r \in \{1, 2, 3, 4\}$ passing held-out energy, directional-balance, spectral-gap and block-jackknife stability checks, or rank zero otherwise (Appendix D). For a symmetric positive-definite request $G \in \mathbb{R}^{r\times r}$ with eigenvalues in $[0.5, 2]$, set

$$C = P_H XU, \qquad \Delta = P_H\{(CG^\top - C)U^\top\}, \qquad X_G = X + \Delta. \tag{3}$$

**Theorem 2** (Certified thin action)**.** *For admissible $U$ and $G$, Equation 3 satisfies*

$$\Delta(I - UU^\top) = 0, \qquad \mathbf{1}^\top\Delta = 0, \qquad P_H X_G U = CG^\top.$$

*Thus the action leaves the full $(K - r)$-dimensional protected complement and the cumulative endpoint unchanged while realizing the requested active response. If $G = I_r$, the implementation returns a bit-exact copy. For any orthogonal $Q$, $(U, G)$ and $(UQ, Q^\top GQ)$ induce the same physical action. The direct compiler costs $O(NHKr + NHr^2)$ time and thin working memory; it never forms a dense $K \times K$ action.*

The construction separates two geometries. Right multiplication by $U^\top$ confines the edit to the active channel subspace, while $P_H$ removes its temporal mean. Since $U^\top U = I_r$ and $P_H^2 = P_H$, projecting the edited increments back onto the active coordinates gives $CG^\top$ exactly. For each path, its centered active second moment therefore changes from $C^\top C/H$ to $G(C^\top C/H)G^\top$. A matrix request can couple active directions; it need not be a scalar rescaling. None of these identities requires a stationary path law. They do require a meaningful active frame.

The frame estimator averages trace-normalized block covariances, then extracts a leading eigenspace. It does not average separate PCA projectors. Held-out energy and directional balance prevent a nominal rank from hiding an inactive direction. The spectral gap and leave-one-block-out projector checks address identification and frame stability. Passing these checks qualifies the declared reference under the tested conditions; it does not show that a future regime will keep the same frame.

**Algorithm 1: Failure-aware KRPC compilation**

1. Validate the request schema. Return a bit-exact NO-OP immediately for $G = I_r$, since an inert request requires no identifiable frame.
2. Qualify the frame for nonidentity requests. Return NO-OP with no active-control claim when $r = 0$; return ABSTAIN when stability or support fails. Validate symmetry, positive definiteness, eigenvalue bounds, and the registered request family.
3. Compute $C$, $\Delta$, and $X_G$ through Equation 3 using only thin products.
4. Check protected-complement, cumulative-endpoint, active-coordinate, and covariance residuals. Return ABSTAIN if any tolerance fails; otherwise return $X_G$, APPLY, and the certificate vector.

The selective-risk policy narrows the compiler's eigenvalue envelope to $[0.75, 1.25]$; feasibility alone does not imply acceptance. It also checks relative action size $a = \|\Delta\|_F/\max(\|X\|_F, \epsilon_{\text{tiny}})$ and support input $d$. In the synthetic study, $d$ is supplied by construction; it is not a validated real-data estimator. Refusal does not license relaxing a deployed threshold. Appendices C and J give the transfer procedure and proofs.

These identities use standard linear algebra; other subspace transforms can share the gates and certificates. Unlike in-sampler projected diffusion (Narasimhan et al., 2025), KRPC exposes a post-composable path action. At rank one, the request scales a single active coordinate. At higher rank, a matrix can couple several active coordinates; a change of basis in $U$ requires conjugating $G$ to keep the physical action unchanged. We claim neither plant controllability nor runtime superiority over constrained samplers.

Table 1: **Which operation is being tested?** Each certificate is tied to its own reference object. These operations share a decision framework, not interchangeable invariants.

| Operation | What stays fixed | Evidence it supplies |
|---|---|---|
| Anchor–flow generator | Anchor ensemble anomalies | EEG joint-generation quality |
| Increment KRPC | Channel complement and cumulative endpoint | Request execution and refusal |
| Paired-slack/event correction | Reference ensemble means, pairs and support; events in later variants | MERRA-2 quality attribution and trade-offs |
| Ensemble-anomaly action | Ensemble mean, pair contrasts and complement; not cumulative endpoint | GHCNh covariance-control attribution |
| Channel-path permutation | Each channel's empirical path distribution | Solar dependence correction; marginal CRPS unchanged |

**Two limits that matter when composing generators.** Preservation always names a reference. A channel-specific centre shift $Y = X+\delta$ preserves ensemble anomalies but changes a pair contrast by $\delta_i - \delta_j$. A later operator that preserves the contrast of $Y$ cannot claim to have preserved the contrast of $X$. Our composition result makes this reference dependence explicit (Proposition 1). Nor are protected diagnostics sufficient support checks. If a history statistic satisfies $T(CQ) = T(C)$ for every orthogonal $Q$, then it also satisfies $T(CR(C)) = T(C)$ for a history-dependent orthogonal action $R(C)$. An invariant-only monitor cannot distinguish those histories, although acting on the history while holding the base law and truth fixed can change proper risk. Proposition 3 formalizes this limit. Additional relational evidence, not more invariant checks alone, is needed to address such failures.

**A compatible generator with an explicit protection identity.** Our EEG generator learns a shared shift while retaining anchor variation. It is one compatible base, not a prerequisite for KRPC.

For synchronized empirical anchors $A \in \mathbb{R}^{M\times H\times K}$ and a learned flow proposal $F$, define

$$Y_m = A_m + \rho(\bar{F} - \bar{A}), \qquad \rho = 0.5. \tag{4}$$

Then $\bar{Y} = \bar{A} + \rho(\bar{F} - \bar{A})$ while $Y_m - \bar{Y} = A_m - \bar{A}$ exactly. The centre moves while empirical anomaly covariance, including cross-time entries, stays fixed. This certifies an ensemble identity, not stationarity, truth-law equality or non-copying.

Conditional flow matching (Lipman et al., 2023) learns the proposal from synchronized anchors (settings in Appendix F). Only its mean enters Equation 4. A direct centre learner could replace the flow; we test protected generation, not an exclusive benefit of that architecture (Section 4.2).

The same task specification accommodates several useful protection sets (Table 1). KRPC supplies the endpoint/complement case; the weather correction preserves ensemble and event properties. Their empirical gains are attributed to those respective operators.

## 4 Evidence: four questions, four distinct tests

The following cases distinguish protected quality, a quality–coverage trade-off, high-rank execution and an adoption decision. They share a task specification, not a fitted generator or a universal protection set. Table 1 maps each operator to its own certificate.

### 4.1 Protected quality on locked annual weather panels

MERRA-2 (Gelaro et al., 2017) provides the main real-data test: can a correction improve generated atmospheric paths while respecting specified reference structure? We use total aerosol extinction optical thickness at 550 nm (TOTEXTTAU) on a fixed 128-cell spatial panel from the hourly aerosol diagnostics product, version 5.12.4 (Global Modeling and Assimilation Office & Pawson, 2015). Each annual panel uses $K = 128, H = 16, M = 64$ and 192 origins. Q0 denotes the proposal, Q1 its calibrated law, and Q2 a protected correction. Each version was frozen before its respective yearly panel was evaluated (Table 2); they are not replications of one unchanged model.

Table 2: **Quality, protected quality, then expert selection.** Locked MERRA-2 evaluations of distinct versions. Reductions use each row's named reference, not the whole roster. Complete contrasts and conditional intervals appear in Appendices M–O.

| Year | Frozen comparison | Measured benefit | Scope |
|---|---|---|---|
| 2023 | Selected Q2 vs own Q1 | Energy $0.426\%$ lower | q99 error $16.73\%$ higher |
| 2024 | Event-preserving path correction vs own Q1 | Six endpoints improve; Energy $0.097\%$, q99 error $0.879\%$ lower | Exceedance identities fixed; Energy rank $8/21$ |
| 2025 | Contextual choice vs mature protected expert | Energy $0.0600\%$ lower; registered checks pass | Target-blind choice; Energy rank $3/22$ |

**Protect the event, then improve the path.** The 2024 multiscale action redistributes paired slack without changing reference ensemble means, within-pair differences, the support floor or member-wise exceedance indicators. Its shape inputs summarize generated paths over 4, 8 and 16 steps. All six endpoints improve over Q1: Energy, marginal CRPS, increment Energy, spatial variogram, q99 amplitude error and path-q99 CRPS. Their 99.8% conditional lower bounds are positive; tail Brier is identical by construction. Crucially, it also improves Energy and increment Energy over the 305-parameter, equal-budget history-conditioned arm. The other four matched contrasts are unresolved. This supports path conditioning beyond the own-Q1 comparison. It is not roster dominance: Q1-anchored correction leads on four metrics (Appendix N).

**Choose a protected expert without seeing the target.** For 2025, a ridge gate fitted on 2020 uses eight calendar and generated-expert diagnostics to choose feasible-cone or mature multiscale correction. Coefficients, experts and inference were frozen before the one-shot panel. Target-blind selection excludes the current verifying path; historical labels are used in fitting. The gate chooses the experts on 141 and 51 origins. Energy improves 0.0600% over the mature expert, with a one-sided 95% block-bootstrap lower bound of 0.0001986 for the absolute reduction. All registered secondary, quarterly, Q1 and certificate checks pass. The deployed two-expert policy therefore improves its reference. This comparison does not isolate the gate from the added expert or test the second-view relation studied below. Two affine arms retain lower Energy; post-hoc event-projected affine controls also trade lower Energy for worse pooled-q99 error (Appendix O.2).

**Why protection belongs in the task.** The earlier 2023 version improved Energy, CRPS and increment Energy by 0.426%, 0.195% and 0.777% over Q1, yet worsened q99 amplitude error by 16.73%. Its two registered questions used the same pair, not independent replications. This motivated the later event-preserving specification. Event identity still does not imply correct tail amplitudes; these remain separate measurements.

The sequence distinguishes three questions that a single leaderboard would conflate. The 2023 comparison asks whether a calibrated generator can gain average quality while losing a tail property. The 2024 comparison asks whether a specified correction can retain event membership and improve quality, including against a matched history-conditioned alternative. The 2025 comparison asks whether a frozen choice between protected experts improves the deployed reference. Its positive answer concerns that complete policy, not every component in isolation. Accordingly, the annual gains should not be added or pooled into a single effect size. The strongest evidence for the mechanism is the matched 2024 contrast; the 2025 result supports transfer of the complete selection policy under its stated checks. Neither establishes that protection is the only way to improve these paths.

### 4.2 Joint-path quality and its coverage trade-off

The EEG generator shifts the anchor ensemble's centre using a learned flow proposal while preserving every member's anomaly about that centre. Equation 4 and Appendix F give the construction. It is a compatible generator, not a requirement for KRPC. EEGMMIDB-R02 (Schalk, 2009) uses S001–S032 for fitting, S033–S048 for trust/development, and S049–S109 as 61 evaluation subjects: eight origins each, $H = 16$, $K = 64$, $M = 32$. The prefix-only sensitivity compares synchronized anchor, one- and 100-step TimeGrad, Diffusion-TS-local, half-mix and our generator. Neural fits use 1,000 updates, batch 16 and context 32 (candidate) or 96 (comparators). These are local config-

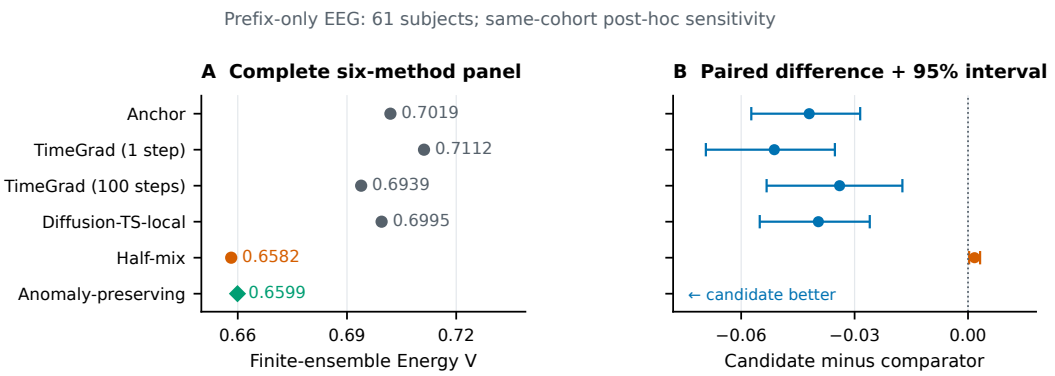


Figure 2: **Generation quality with the stronger EEG controls visible.** Prefix-only post-hoc sensitivity, 61 subjects with eight origins each. A: exact Energy of the issued finite ensemble, lower is better. B: candidate minus each comparator, with marginal 95% paired-subject bootstrap intervals conditional on the saved fits and sampling tapes. Half-mix has lower Energy than the protected candidate. These are not simultaneous intervals, new promotion tests or independent confirmation.

urations, not optimized published reproductions. Prior use of the same cohort makes this post-hoc sensitivity, not fresh confirmation.

Because the mean shift couples ensemble members, our principal display uses finite-ensemble Energy V, not an unbiased marginal-law interpretation of the historical off-diagonal score. The candidate's 0.659923 versus 100-step TimeGrad's 0.693891 is a 4.895% point reduction. Its paired difference is $-0.033968$ with interval $[-0.053272, -0.017358]$. The interval uses saved subject aggregates and the original 20,000 bootstrap addresses, not an off-diagonal-score interval. All five comparisons and both units appear in Appendix F.2.1.

The strongest mechanistic alternative matters. Half-mix has lower V-score (0.658202), but larger coverage error (0.084633 versus 0.070117). The protected candidate also has lower prefix DCT-CRPS and variogram score, while half-mix leads several physical-unit and marginal comparisons. This is a trade-off, not dominance. Zero exact copies and positive nearest-training distances are narrow non-copy evidence, not privacy. Original gates and future-dependent normalization remain in Appendix D; all sensitivity results remain in Appendices F.1–F.2.

### 4.3 High-rank request execution on 500 channels

On fixed D3U streamflow ensembles, a rank-4 request-conditioned native operator lowers normalized request error by 90.36% versus identity and 91.07% versus a parameter-matched arm without request geometry. The 128 unseen context/request cells form a later fit-internal partition; all thin lifts retain the complement and endpoint (Table 3). This tests learned execution on 500 hydrological stations. Separately, the nine-method USGS quality panel credits D3U with its 3.6366% Energy lead (U.S. Geological Survey, 2026; Li et al., 2025), not the controller.

This study learns an approximation to the analytic thin-coordinate target. Request error measures that approximation; the subsequent thin lift preserves the complement and endpoint independently of that error. The two learned arms each have 20,999 parameters and receive 5,000 updates on the same 32 early contexts, saved paths, operator targets, batches and initialization. Eight later contexts with 16 disjoint requests each define the operator test. The contrast isolates access to request geometry within this fitting setup. It does not show that learning is preferable to applying the analytic compiler when the requested matrix is already available.

**Qualification and refusal have separate scopes.** The fixed-rank atlas uses $r = 4$ at $K = 32$; its $K = 128$ bridge extends the complement spectrum and rescales active eigenvalues by the new-to-old complement-trace ratio. Its simultaneous bounds, 0.98946 for valid acceptance and 0.01054 for false acceptance, do not establish arbitrary dimension transfer (Appendix D.2). The separate adaptive-rank test at $K = 32$ initially recovers rank four in only 169/200 cases. A development-

Table 3: USGS $K = 500$ fit-only operator-control mechanism result over 128 unseen context–request cells. Lower normalized request error is better. Identity and the matched arm do not receive extra model capacity.

| Arm | Mean normalized error | 95th percentile |
|---|---|---|
| Identity | 1.000000 | – |
| Matched unrequested | 1.079478 | – |
| Requested native fiber | 0.096359 | 0.357096 |

Table 4: **A relation changes the decision, not the guarantee's scope.** All five registered scenarios of the same synthetic study. Apply counts refer to 32 frozen policy pairs, not individual paths. Energy change is $100(\overline{S}_{\mathrm{deployed}}/\overline{S}_{\mathrm{scalar}} - 1)$; negative is better. A zero denotes exact fallback, not a newly improved generator. These are descriptive counts and means, not simultaneous safety bounds.

| | Original screen | | With relation | |
|---|---|---|---|---|
| Past-view condition | Apply / 32 | Energy, % | Apply / 32 | Energy, % |
| Clean | 32 | $-0.1208$ | 32 | $-0.1208$ |
| Moderate auxiliary noise | 32 | $-0.1208$ | 32 | $-0.1208$ |
| Severe primary noise | 0 | 0.0000 | 0 | 0.0000 |
| Primary-only reversal | 32 | $+0.1060$ | 0 | 0.0000 |
| Reversal of both views | 32 | $+0.1060$ | 32 | $+0.1060$ |

selected sample-size increase reaches 197/200 on fresh synthetic draws (Appendix D.3). Across 5,400 action/refusal cases at $K = 32, 128$ and $r = 1, 2, 4$, safe coverage is 0.9928, with no invalid request applied in 1,800 trials and 1,800 bit-exact identity/zero-rank returns. These are named-family checks, not a universal guarantee (Figure 5).

### 4.4 When does additional information change the decision?

To isolate the role of information, the frozen synthetic study in Appendix AA uses 32 previously fitted scalar/directional policy pairs, $M = 64$, $H = 16$ and 512 evaluation origins per pair. The learned request may change active rank-4 geometry; the base path's complement and endpoint must remain fixed. Its primary quality objective is fair Energy relative to the scalar policy, not action error alone. Scores use the $16 \times 4$ thin paths; the $K = 500$ lifts check geometry. Two 96-step past views share latent history but have independent observation noise. The rule uses past-derived shape estimates, not future values, true shapes or scores.

The original scale/agreement screen accepts a primary-view spectral reversal in 32/32 policy pairs, worsening Energy by 0.1060%. Adding the two views' normalized shape alignment rejects this case. The frozen rule requires validation activation, the old screen and a batch-median alignment drop no larger than five training-reference median absolute deviations. Its threshold comes from disjoint development. One decision covers all 512 origins; rejection returns scalar paths bit-exactly without refitting.

The additional relation changes the primary-reversal decision without sacrificing clean operation (Table 4). Clean gains range from 0.0748% to 0.1782% across policy pairs. Shared reversal still harms every pair, by 0.0559–0.1611%; these ranges are descriptive, not confidence bounds. Eight $K = 500$ lifts pass the geometric checks, not a per-origin risk guarantee. The full law, inherited fits, thresholds and a saved-record replay are specified in Appendix AA.

The decision table keeps rejected cases in the outcome denominator. Always returning the scalar policy would remove the reversal losses, but would also forgo the clean-case improvement. Always using the directional policy would retain that gain while accepting unsupported changes. The additional relation separates these choices only when the two views disagree informatively. If both views undergo the same reversal, agreement is restored even though the relation to the verifying law is wrong. This is why exact geometric protection, a past-only input rule and observed cross-view agreement are each insufficient for unconditional benefit. The result identifies information needed for adoption; it does not establish how frequently the five scenarios occur in an application.

## 5 Discussion: what the evidence does and does not establish

**Selective fidelity is not a fixed global covariance.** A protection reference may change with context. The distinct annual weather versions show scoped quality gains, not unchanged-model replications. EEG and tail-focused development illustrate permitted costs: the latter improves its primary tail score at a small Energy cost (Appendix O.4). Success need not mean winning every secondary metric.

**How to use the specification.** A new application should first state the desired change in observable task terms: for example, alter joint fluctuations while retaining cumulative totals and a named channel contrast. This defines an editable direction and a protected reference without requiring a particular neural architecture. Next, it should establish which evidence can evaluate the action. A request-matched simulator can supply a quality target; an algebraic request alone cannot. When only ordinary observations are available, base-law quality and action execution remain separate questions.

Finally, the application should compare complete decision rules, not only accepted outputs. Direct projection, a simple calibrated adjustment and always returning the base are useful controls whenever they meet the same task definition. Report the primary outcome over all eligible cases, together with protection violations, refusal, diversity and allowed costs. A changing environment may justify a new reference, but that reference must be selected without using the current verifying outcome. This workflow allows adaptation without treating arbitrary changes as realistic or equating an exact certificate with a learned model of regime evolution.

**The interface is reusable; its executor need not be unique.** SVD projection can reproduce graph protection; protected whiten–color and dense transport tie KRPC on Solar. Feature Energy also matches or exceeds our dynamic moment statistic (Appendix G). The contribution is the reference-aware separation of quality, action and adoption with a thin realization, not exclusive ownership of projection or proper-score training (Pacchiardi et al., 2024).

**Remaining limits.** High-rank real-use benefit and adaptation to unknown nonstationarity remain open. EEG uses local, limited-budget baselines and a reused cohort. Shared distortion defeats the synthetic relation check; the 2025 gain does not isolate the selector from its added expert. Certificates imply neither causal benefit nor privacy. Temporal dependence, coupled-ensemble scoring, physical support issues and other adverse results are documented in Appendices B, G and Q.

## 6 Conclusion

Selective structural fidelity separates plausible generation, correct action and justified adoption. Our reference-aware contract and thin controller make these claims testable. Matched controls and locked-panel evaluations support scoped gains; the counterexamples show why exact protection alone cannot establish benefit.

### Reproducibility and data availability statement

Per-study protocols and saved result records document the frozen evaluators and thin compiler. EEGMMIDB, PTB-XL, and NOAA OISST are public under their provider terms; USGS NWIS daily values are likewise public. The MERRA-2 analysis uses NASA Earthdata under its access terms. Root-law selection uses 2022 and the original confirmation uses 2023. The separately locked event-preserving comparison uses 2024. The 2021 split belongs to historical development; it is not fresh for the new method. Only 2025 was reserved for the final contextual-gate one-shot and is now consumed. Raw real data are not redistributed. The manuscript-source archive supports typesetting only. A separate central-case package contains the original synthetic source, inherited coefficients, saved records and a read-only arithmetic replay for Table 4 (Appendix AA). It is not a new scientific replication or whole-paper replay. Other studies' data, terminal weights and third-party sources remain separate dependencies.

### Ethics statement

The study analyzes public, de-identified physiological and environmental data, recruits no participants, and makes no clinical or physical-intervention claim.

### AI use statement

The human authors led this study and directed the drafting and revision of all sections. They set the research questions and conceptual direction and made the final decisions on algorithmic design, experimental protocols, analysis, and conclusions. Generative AI tools, including OpenAI Codex, provided assistance under their direction in five areas: writing support and polishing; retrieval and discovery of related work; research ideation and execution; drafting sections of the paper; and proving mathematical claims. Research assistance included hypothesis refinement, methodological critique, experimental design, data processing, implementation and debugging, result interpretation, citation organization, and figure planning. Mathematical assistance included formulating claims and developing, writing, and checking proofs.

AI tools had no independent authority to open held-out targets, change frozen thresholds, designate an evidence class, or approve a scientific claim. The authors are responsible for reviewing and validating AI-assisted work and for all final text, proofs, code, citations, numerical results, and interpretations.

## References


Sam Allen, David Ginsbourger, and Johanna Ziegel. Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):906–940, 2023. doi: 10.1137/22M1532184. URL https://epubs.siam.org/doi/10.1137/22M1532184.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022. URL https://arxiv.org/abs/2110.01052. arXiv:2110.01052, version 5.

Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Nicolas Chapados, and Alexandre Drouin. TACTiS-2: Better, faster, simpler attentional copulas for multivariate time series. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01327.

Andrea Coletta, Sriram Gopalakrishnan, Daniel Borrajo, and Svitlana Vyetrenko. On the constrained time-series generation problem. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper/2023/hash/bfb6a69c0d9e2bc596e1cd31f16fcdde-Abstract-Conference.html.

Xilin Dai, Zhijian Xu, Wanxu Cai, and Qiang Xu. From samples to scenarios: A new paradigm for probabilistic forecasting. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.19975.

Alexandre Drouin, Étienne Marcotte, and Nicolas Chapados. TACTiS: Transformer-attentional copulas for time series. In *Proceedings of the 39th International Conference on Machine Learning*,

volume 162 of *Proceedings of Machine Learning Research*, pp. 5447–5493. PMLR, 2022. URL https://proceedings.mlr.press/v162/drouin22a.html.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270. URL https://doi.org/10.1002/qj.2270.

Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017. doi: 10.1175/JCLI-D-16-0758.1. URL https://doi.org/10.1175/JCLI-D-16-0758.1.

Global Modeling and Assimilation Office and Steven Pawson. MERRA-2 tavg1_2d_aer_nx: 2d,1-hourly,time-averaged,single-level,assimilation,aerosol diagnostics V5.12.4, 2015. URL https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4/summary.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. *arXiv preprint arXiv:2505.04163*, 2025. doi: 10.48550/arXiv.2505.04163. URL https://arxiv.org/abs/2505.04163.

Boyin Huang, Chunying Liu, Viva F. Banzon, Eric Freeman, Garrett Graham, William Hankins, Thomas M. Smith, and Huai-Min Zhang. NOAA 0.25-degree daily optimum interpolation sea surface temperature (OISST), version 2.1, 2020. URL https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C01606/html. Subset used is declared in the manuscript.

Kelvin Kan, François-Xavier Aubet, Tim Januschowski, Youngsuk Park, Konstantinos Benidis, Lars Ruthotto, and Jan Gasthaus. Multivariate quantile function forecaster. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10603–10621. PMLR, 2022. URL https://proceedings.mlr.press/v151/kan22a.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*. OpenReview.net, 2025. URL https://openreview.net/forum?id=HdUkF1Qk7g.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Étienne Marcotte, Valentina Zantedeschi, Alexandre Drouin, and Nicolas Chapados. Regions of reliability in the evaluation of multivariate probabilistic forecasts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23958–24004. PMLR, 2023. URL https://proceedings.mlr.press/v202/marcotte23a.html.

Sai Shankar Narasimhan, Shubhankar Agarwal, Litu Rout, Sanjay Shakkottai, and Sandeep Chinchali. Constrained posterior sampling: Time series generation with

hard constraints. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107923–107963. Curran Associates, Inc., 2025. doi: 10.52202/085713-3599. URL https://papers.nips.cc/paper_files/paper/2025/hash/9b01c4a7d3fc49875dad3c13848bcd9e-Abstract-Conference.html.

NOAA National Centers for Environmental Information. Global historical climatology network hourly (GHCNh), 2026. URL https://www.ncei.noaa.gov/products/global-historical-climatology-network-hourly. Version 1.1.0 product documentation; accessed for the frozen paired-station study.

Lorenzo Pacchiardi, Rilwan A. Adewoyin, Peter Dueben, and Ritabrata Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024. URL https://www.jmlr.org/papers/v25/23-0038.html.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1):23–58, 2025. doi: 10.5194/ascmo-11-23-2025. URL https://ascmo.copernicus.org/articles/11/23/2025/.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Gerwin Schalk. EEG Motor Movement/Imagery Dataset. *PhysioNet*, 2009. doi: 10.13026/C28G6P. URL https://physionet.org/content/eegmmidb/1.0.0/. Version 1.0.0.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072. URL https://pubmed.ncbi.nlm.nih.gov/15188875/.

Roman Schefzik. Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quarterly Journal of the Royal Meteorological Society*, 142(699):2463–2477, 2016. doi: 10.1002/qj.2839. URL https://doi.org/10.1002/qj.2839.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1. URL https://doi.org/10.1175/MWR-D-14-00269.1.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31016–31029. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23d.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011. doi: 10.18910/4973. URL https://doi.org/10.18910/4973.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

U.S. Geological Survey. National water information system: Daily values, 2026. URL https://doi.org/10.5066/F7P55KJN. Accessed August 2026.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset.

*Scientific Data*, 7(1):154, 2020. doi: 10.1038/s41597-020-0495-6. URL https://doi.org/10.1038/s41597-020-0495-6.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, 2022. doi: 10.13026/kfzx-aw45. URL https://physionet.org/content/ptb-xl/1.0.3/. Version 1.0.3.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. URL https://doi.org/10.1093/biomet/asv008.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, pp. 41582–41610, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b5b66077d016c037576cc56a82f97f66-Abstract-Conference.html.

Jiawen Zhang, Xumeng Wen, Zhenwei Zhang, Shun Zheng, Jia Li, and Jiang Bian. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. In *Advances in Neural Information Processing Systems*, volume 37, pp. 48045–48082. Curran Associates, Inc., 2024. doi: 10.52202/079017-1523. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/55f2a27b1ac39dbfdd0fc83742dc87d7-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Shiqiao Zhou, Holger Schöner, Zipeng Wu, Edouard Fouché, IAG Wilson, and Shuo Wang. Stationarity-aware retrieval-augmented time series forecasting. *arXiv preprint arXiv:2606.04135*, 2026. doi: 10.48550/arXiv.2606.04135. URL https://arxiv.org/abs/2606.04135. Accepted to the KDD 2026 research track according to the arXiv record.

**Appendix reading guide.** For the main claims, see the Energy interpretation and proof (Appendices B and H), the KRPC guarantees and qualification (Appendices J and D), the annual weather comparisons (Appendices M–O), the EEG sensitivity (Appendix F.2.1), and the central decision replay (Appendix AA). The remaining domain studies document attribution and limits; they are not additional confirmations of one unchanged model.

## A Closest work and scalar/matrix request geometry

RAFT and SARAF already establish retrieval-augmented forecasting (Han et al., 2025; Zhou et al., 2026); retrieval alone is therefore not novel. A scalar request is a special case of Equation 3. For $K = 2$, choosing $U = (1,1)^\top/\sqrt{2}$ edits common motion while its orthogonal complement $(1,-1)^\top/\sqrt{2}$ protects the paired contrast. For general $K$ and rank $r$, the protected complement has dimension $K - r$. Matrix requests may couple active coordinates, and basis changes require conjugating the request. The guarantee follows from the same displayed identity; increasing the channel count alone is not the contribution.

Pic et al. (2025) formalize proper scoring after transformations and nonnegative aggregation. A lower-dimensional transformation may target a useful feature without distinguishing every path law. Our dynamic comparison therefore includes Energy on the same features, rather than attributing its representation advantage to a uniquely new score. The added object here is the reference-bound action and adoption interface: a proper feature score does not by itself certify a protected coordinate, detect an invisible history distortion, or justify the requested law.

## B Finite predictive measures and outer-unit inference

**Which risk is being estimated?** Fix an observed history $c$ and all fitted artifacts. Let $Y \sim P$ be independent of the ensemble $Z_{1:M}$ conditional on this information, $M \geq 2$. Suppose $P$ and the common member marginal law $Q$ both have finite first moments, without assuming independent members. Define

$$D_Q = \mathbb{E}\|Z - Z'\|_2 \quad (Z, Z' \overset{\text{iid}}{\sim} Q), \qquad D_{\text{pair}} = \frac{1}{M(M-1)} \sum_{m \neq n} \mathbb{E}\|Z_m - Z_n\|_2,$$

Table 5: Closest-work boundary. Prior capability is stated before the remaining distinction; the table does not claim universal priority.

| Family | Established capability | Remaining distinction here |
|---|---|---|
| TimeGrad/TimeDiff | Diffusion-based generation of future trajectories | Failure-qualified promotion and certified path action |
| CSDI | Conditional generation for time-series imputation | Disclosed future-block adaptation, not a new imputation model |
| TACTiS/MQF2 | Joint predictive distributions with likelihood or proper-score training | Outer-unit inference and ordered veto gates |
| Diffusion-TS | General time-series generation | No backbone novelty claim; disclosed local adaptation |
| Sundial | A generative foundation model for forecasting | No foundation-scale or pretraining novelty claim |
| COP / GuidedDiff-Time | Post-hoc constrained adjustment of another generator, and constraint-guided diffusion (Coletta et al., 2023) | Specified thin matrix response, complement/endpoint certificate and fixed refusal semantics; not first post-processing |
| Learn then Test | Model-agnostic calibration and simultaneous risk control under its assumptions (Angelopoulos et al., 2022) | Reference-aware path geometry and named-failure tests; not a new general risk-control theorem |
| RAFT/SARAF | RAFT retrieves historical examples; SARAF adds stationarity-aware retrieval and diversity selection | Retrieval is a source; calibration and non-copy generation remain gated |
| Regions of Reliability | Finite-sample power limits of proper scores (Marcotte et al., 2023) | Named corruptions and operational promotion rules |
| CPS | Hard constraints through projected diffusion sampling | Thin post-composable action, protected complement, endpoint, and abstention certificates |

and $R(Q;P) = \mathbb{E}_{Y\sim P} S_{\mathrm{E}}(Q,Y)$. The score of the issued measure $\widehat{Q}_Z = M^{-1}\sum_m \delta_{Z_m}$ is exactly $S_{\mathrm{V}}$ in Equation 19, whatever the coupling. Taking expectations of its two terms, and of the historical off-diagonal score, gives

$$\mathbb{E}S_{\mathrm{U}} - R(Q;P) = \tfrac{1}{2}(D_Q - D_{\mathrm{pair}}), \tag{5}$$

$$\mathbb{E}S_{\mathrm{V}} - R(Q;P) = \tfrac{1}{2}\{D_Q - (1-M^{-1})D_{\mathrm{pair}}\}. \tag{6}$$

The proof is direct: both first terms have expectation $\mathbb{E}_{Z\sim Q, Y\sim P}\|Z-Y\|_2$; the population, U, and V dispersion terms are respectively $D_Q/2$, $D_{\mathrm{pair}}/2$, and $(1-M^{-1})D_{\mathrm{pair}}/2$. For iid members, U is unbiased and V has bias $D_Q/(2M)$ relative to the marginal-law risk. For dependent members, neither conclusion follows from exchangeability alone, consistent with Ferro (2014). If member marginals differ, these common-$Q$ identities do not apply, but the issued-measure identity remains exact.

The newer protocol therefore estimates $\mathbb{E}_{Z,Y} S_{\mathrm{E}}(\widehat{Q}_Z, Y)$, the risk of the random finite path distribution actually issued. Scoring $\widehat{Q}_Z$ is exact within that finite measure. Averaging these scores over verification cases still has sampling uncertainty and is not generally unbiased for the latent member-law risk. Common $M$ prevents a sample-count mismatch. It does not remove coupling bias or ensure sufficient sensitivity to dependence. If every $Z_m$ equals the same random path $Z$, increasing $M$ leaves $S_{\mathrm{V}} = \|Z-Y\|_2$ unchanged: no universal $1/M$ variance reduction or member-based effective sample size is available. The exchangeable two-member counterexample in Appendix K gives the opposite, negatively coupled case. These identities do not change any historical scoring convention or promotion decision.

**Conditional uncertainty and its limits.** Let $\mathcal{C}$ denote the frozen fits, selection decisions, normalization and the random artifacts on which an analysis conditions. Within outer unit $j$, form the prescribed paired score contrast $D_j$, after reducing all of its origins, members and other repeated

axes. For $\bar{D} = n^{-1} \sum_j D_j$ and finite second moments, the exact identity is

$$\mathrm{Var}(\bar{D} \mid \mathcal{C}) = \frac{1}{n^2} \left[ \sum_j \mathrm{Var}(D_j \mid \mathcal{C}) + 2 \sum_{j<k} \mathrm{Cov}(D_j, D_k \mid \mathcal{C}) \right]. \tag{7}$$

If, additionally, the outer contrasts are conditionally iid with mean $\mu_{\mathcal{C}}$ and nonzero finite variance $\sigma^2_{\mathcal{C}}$, then this variance is $\sigma^2_{\mathcal{C}}/n$ and the usual central limit theorem applies. A paired outer bootstrap preserves all within-unit dependence. Its usual asymptotic interpretation requires the corresponding sampling and regularity assumptions; finite-sample coverage is not implied by either the Energy identity or geometric certificates. Freezing a fit or tape does not by itself establish iid outer units. If an unconditioned random source is shared across units, its induced covariance must remain in Equation 7.

The mean of subject-relative reductions and the reduction of pooled means are distinct statistics. For the latter, $g = 1 - \mu_A/\mu_B$, with $\mu_B > 0$, the first-order outer influence is $-(A_j - \mu_A)/\mu_B + \mu_A(B_j - \mu_B)/\mu_B^2$. The ratio requires joint finite second moments, a denominator separated from zero, and the applicable outer CLT/bootstrap assumptions. Each study retains its originally specified estimand and resampling statistic. Descriptive ratios without registered intervals are not upgraded to ratio inference.

Table 6: Study-specific units for uncertainty estimation. Each row gives the declared grouping and its remaining assumptions; the grouping alone does not establish independence.

| Study | Reduction and inferential grouping | Remaining uncertainty boundary |
|---|---|---|
| Known-truth requests | One generated frame/request replicate | Exact binomial bounds for registered truth families; multiplicity fixed |
| EEG R02 | Eight origins per subject; 61 paired subjects | Conditional on one fit and retained tapes; post-hoc prefix cohort is reused |
| USGS | Eight origins per month; 48 month blocks | Residual adjacent-month dependence is not explicitly modeled |
| MERRA-2 locked years | Replicates reduced within origin; chronological circular blocks | Block length 16 is frozen; finite seasonal sequence has no distribution-free coverage guarantee |

Subject intervals do not integrate training-seed or selection uncertainty. Chronological block intervals require further dependence assumptions, and Bonferroni adjustment cannot repair invalid component coverage. The exact binomial qualification bounds and conditional real-data bootstrap intervals therefore have different guarantees.

## C TRANSFER, CALIBRATION, AND REFUSAL BOUNDARIES

Table 7 separates the reusable interface from its instantiations. In the known-truth selective panel, $d$ is an externally supplied dimensionless support input with boundary one and safe value 0.4 by construction. It is not an estimated EEG or MERRA-2 support probability. The relative action norm $a$ is computable from the proposed paths, but its 0.20 limit and the gain envelope are registered policy choices, not distribution-free safety constants.

For a new domain, first define the protected object, outer unit, support representation and candidate policy using FIT information. Choose tolerances on a separate qualification set, then freeze the entire policy before evaluating untouched validation data. If several candidate policies are certified on the same validation set, simultaneous bounds must cover all of them before selection, or a fresh validation set must assess the selected policy. This procedure is guidance for a future study, not a claim that a real-data support estimator was validated here. Bounds for registered synthetic failure families do not certify failures outside those families or arbitrary domain shift. Frequent refusal calls for retaining the base output and revisiting representation on development data, not relaxing the deployed cutoff after seeing failures.

The real-data control studies do not sample the known-truth panel's request mixture. EEG evaluates three fixed scalar probes; the prefix sensitivity checks its declared fixed-rank certificates, not a new

rank-selection qualification. MERRA-2 corrections and expert selection use different feasible sets. Certificate pass rates or expert-selection counts cannot be relabelled as a deployment-wide ABSTAIN frequency. Such a rate needs a specified incoming-request distribution and retained denominators, which this paper does not estimate for those real-data applications.

Existing checks cover ranks 1, 2 and 4, named spectral and frame failures, noisy-context qualification and a shared-bias counterexample (Appendix Q). They do not establish a real-data rank/noise robustness curve. Likewise, the frozen MERRA-2 diagnostics include increment Energy, spatial variograms, path-q99 CRPS and tail Brier scores. Their disagreements with full-path Energy are reported, including tail harm in 2023. A systematic graph/localized-score comparison remains unperformed; adding one now would be post-hoc and would not replace the registered primary endpoint. A prospective primary endpoint must be qualified and fixed before new verification data are consumed.

The comparison with constrained posterior sampling (CPS) concerns architecture. Empirical success rates and runtimes have not been compared. A shared benchmark would need the same base sampler, information, admissible requests, invariants and compute-accounting rules before either method is run. Single-seed budget-limited fits and the absence of that head-to-head remain evidence gaps. A schedule-only TimeGrad sensitivity would not, by itself, establish multi-seed stability, convergence or compute-matched superiority.

**Available replay and missing reproduction layers.** The local EEG prefix statistical-replay artifact contains the completed origin scores, fixed provenance and an independent NumPy implementation of the original subject aggregation and bootstrap. It reproduces aggregate statistics from saved scores. It does not recompute scores from raw EDF targets, reproduce model training, or rerun all qualification and locked-year experiments. The typesetting archive and scoped code handoff have still different purposes. Public distribution, third-party licensing and complete scientific reproduction remain separate release tasks.

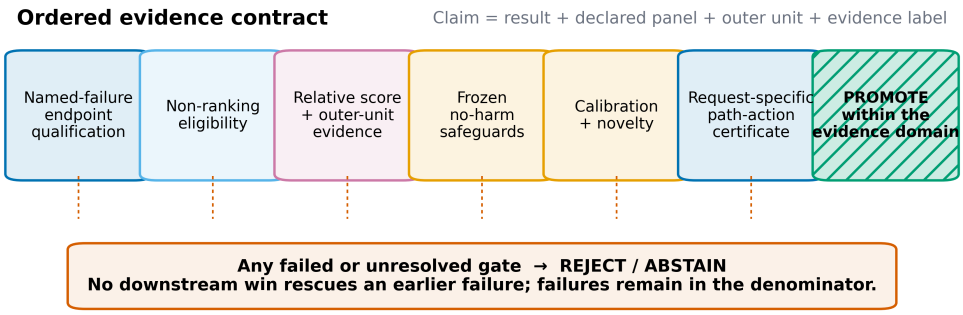


Figure 3: The ordered promotion contract. A downstream advantage cannot repair an upstream failure; every final claim carries its evidence domain.

## D ENDPOINT QUALIFICATION AND COMPLETE GATE REGISTRY

Promotion is the ordered conjunction

$$\mathcal{P} = \mathcal{E}_{\text{source}} \cap \mathcal{E}_{\text{eligibility}} \cap \mathcal{E}_{\text{relative}} \cap \mathcal{E}_{\text{no harm}} \cap \mathcal{E}_{\text{calibration}} \cap \mathcal{E}_{\text{novelty}} \cap \mathcal{E}_{\text{control}}. \tag{8}$$

A gate is a required pass/fail check; an endpoint is its measurement. The contract distinguishes exact algebraic protection, distribution-specific measurement qualification and statistical evidence for adoption. Named corruptions and thresholds are fixed before prospective ranking. Our structural challenges qualify specific measurements (Appendices X and Z); not every EEG quality or non-copying endpoint has an independent power study. *Certified promotion* therefore means acceptance under the stated, tested contract, not universal safety. The assurance and transfer boundaries appear in Table 7.

We distinguish a measurement's ability to detect a named failure from a generator's passing a numerical threshold. The frame atlas below qualifies the former for a specified synthetic distribution. The request study in Appendix E separately tests action and refusal. Table 10 records the historical EEG generator's gates; it does not establish a matched-corruption power study for each of its quality, calibration and non-copying endpoints. These are different evidence levels.

Table 7: What transfers is the decision contract, not universal thresholds. The right-hand column lists declarations required for prospective use; adaptive studies remain labelled as such.

| Component | Reusable requirement | Study-specific declaration |
|---|---|---|
| Source | Freeze information and fitted artifacts | FIT data, preprocessing, budget, seed |
| Inference | Reduce within units before paired inference | Subject/block grouping, weights, dependence model |
| Quality | Qualify endpoints, then apply non-compensating gates | Corruptions, score object, tolerances, multiplicity |
| Non-copying | Test retained paths, not only mean scores | Reference bank, distance scale and cutoff |
| Control | Validate requests and return certificates or refusal | Frame, support input, gain/action limits, invariants |

### D.1 Frame estimand, held-out statistics and refusal

For standardized level paths $Y_{bi} \in \mathbb{R}^{(H+1)\times K}$ in block $b$, define centered increments and trace-normalized block covariance by

$$D_{bi} = P_H \Delta Y_{bi}, \qquad C_b = (N_b H)^{-1} \sum_{i=1}^{N_b} D_{bi}^\top D_{bi}, \qquad S_b = C_b / \operatorname{tr}(C_b), \tag{9}$$

$$A_I = |I|^{-1} \sum_{b \in I} S_b, \qquad U_I = \operatorname{eig}_{1:r}(A_I), \qquad P_I = U_I U_I^\top. \tag{10}$$

The estimand is the top-$r$ eigenspace of the equally weighted normalized block covariances, not an average of block PCA projectors. With eigenvalues in descending order and $-b$ denoting all blocks except $b$, the statistics are

$$e_b = \operatorname{tr}(U_{-b}^\top S_b U_{-b}), \qquad a_b = \frac{e_b}{\sum_{j=1}^r \lambda_j(S_b)}, \qquad v_b = \frac{r\lambda_{\min}(U_{-b}^\top S_b U_{-b})}{e_b}, \tag{11}$$

$$g_I = \frac{\lambda_r(A_I) - \lambda_{r+1}(A_I)}{\lambda_1(A_I)}, \quad \rho_J = \max_{b<c} \|P_{-b} - P_{-c}\|_2, \quad \rho_F = \max_b \|P_{\text{all}} - P_{-b}\|_2. \tag{12}$$

The six-vector is $(\min_b e_b, \min_b a_b, \min_b v_b, \min\{g_{\text{all}}, \min_b g_{-b}\}, \rho_J, \rho_F)$. Projector distance is spectral norm, the sine of the largest principal angle, not Frobenius norm. Each gate in Table 8 must pass; good average performance cannot compensate for a failing block.

Table 8: Frozen scientific frame thresholds and their calibration sentinels. These thresholds are distinct from numerical solver tolerances.

| Statistic | Threshold | Rule | Designated negative family |
|---|---|---|---|
| $\min e_b$ | 0.20662646122223746 | $\geq$ | Low common energy |
| $\min a_b$ | 0.6786426198332047 | $\geq$ | Local-oracle takeover |
| $\min v_b$ | 0.3925317776810288 | $\geq$ | Single-direction dominance |
| $\min g_I$ | 0.1901115251390724 | $\geq$ | Near-flat identified boundary |
| $\rho_J$ | 0.13380549582764528 | $\leq$ | Gradual frame drift |
| $\rho_F$ | 0.5194334007597606 | $\leq$ | Single influential block |

The fixed-rank implicit solver separately rejects degenerate trace, insufficient rank, an unidentified pooled or leave-one-block-out boundary, and excessive eigen-residual. Its relative rank, boundary and residual tolerances are $10^{-10}$, $10^{-8}$ and $10^{-7}$; ARPACK uses tolerance $10^{-10}$, 2,000 iterations, oversampling 12 and chunk size 32. Block-oracle boundary separation is not an additional scientific gate. Numerical diagnostics guard denominators by Float64 tiny and clip energy ratios to $[0, 1]$ only for floating-point roundoff.

### D.2 CONTINUOUS-TIMELINE ATLAS AND THRESHOLD SELECTION

The original atlas fixes $K = 32, r = 4, H = 16, L = 96, B = 8$. Each replicate is one 7,392-row hourly level timeline: a 672-row nuisance prefix, a 168-row embargo, then eight 672-row geometry cores separated by seven further 168-row embargos. Channel means and population SDs are fitted once on the prefix. For a core starting at row $s_b$, its 36 origins are $s_b + 95 + 16j$, $j = 0, \dots, 35$; each extracted level path has 17 rows. Both its 96-row context and future remain in that core. This experiment does not remove the across-channel mean and does not substitute independent path tensors for the continuous timeline.

A Gaussian QR with positive diagonal convention gives $Q \in O(32)$; let $U = Q_{:,1:4}$ and $V = Q_{:,5:8}$. Unless stated otherwise below, a core has standardized increment covariance

$$\Sigma_b^{\mathrm{std}} = cI + U_b \operatorname{diag}(\ell - c\mathbf{1})U_b^\top. \tag{13}$$

Prefix increments are iid Gaussian with covariance $I + U \operatorname{diag}(5, 4, 3, 2)U^\top$. If $D$ contains the SDs of the integrated prefix levels, core covariance in raw units is $D\Sigma_b^{\mathrm{std}}D$. Embargos use the prefix covariance. Following the prefix, the increment recursion starts at zero and obeys $x_t = \phi x_{t-1} + \sqrt{1-\phi^2}L_t\epsilon_t$, with $L_t$ the asset-order lower Cholesky factor of the raw covariance; levels are cumulative sums of increments. Innovations have independent standard-normal coordinates, except the heavy-tail family, which uses $\sqrt{5/7}\, t_7$. Prefix innovations remain Gaussian. For each core, let $V_{b,0}$ be its entry-state increment covariance in prefix-standardized coordinates. At the first used future, after $L + 1 = 97$ updates within that core, the relative transient bound $\phi^{194}\|V_{b,0} - \Sigma_b^{\mathrm{std}}\|_2/\lambda_{\min}(\Sigma_b^{\mathrm{std}})$ must be at most $10^{-12}$. It is evaluated in the log domain, with value zero when $\phi = 0$ or the entry and target covariances coincide.

Table 9: Complete fixed-rank atlas families. The first four are valid; the remaining six are named failures. All block indices start at zero.

| Family | Active eigenvalues $\ell$ | $c$ | $\phi$ | Additional rule |
|---|---|---|---|---|
| Stable Gaussian | $(10, 9, 8, 7)$ | .70 | 0 | Common $U$ |
| Stable heavy-tail | $(10, 9, 8, 7)$ | .70 | .65 | Unit-variance $t_7$ |
| Changing complement | $(9, 8, 7, 6)$ before factors | .75 | .25 | Rule below |
| Minimum margin | $(4, 3.7, 3.4, 3.1)$ | 1.20 | .35 | Common $U$ |
| Low energy | $(1.20, 1.18, 1.16, 1.14)$ | 1.00 | .20 | Common $U$ |
| Local takeover | $(4, 3.8, 3.6, 3.4)$ | .40 | .20 | Extra local spike |
| Single direction | $(14, .72, .68, .64)$ | .55 | .20 | Common $U$ |
| Near flat | $(1.0020, 1.0018,$ $1.0016, 1.0014)$ | 1.0010 | 0 | Identified but small gap |
| Frame drift | $(10, 9, 8, 7)$ | .60 | .15 | Rotating $U_b$ |
| Influential block | $(1.05, 1.04, 1.03, 1.02)$ | 1.00 | .15 | Block 7 replaced |

For changing complement, rotate the first two active columns by $.31b$ radians and use active eigenvalues $(.8{+}.08b)(9, 8, 7, 6)$. Along $Q_{:,5:32}$, use the 28 eigenvalues $c_b \operatorname{roll}_{3b}(\operatorname{linspace}(.65, 1.35, 28))$, where $c_b = .75[.75 + .5(b + 1)/8]$. Multiply the full covariance by the respective factor in $(.55, 1.65, .8, 1.35, .7, 1.9, 1.1, .95)$. For local takeover, $W_0, \dots, W_6$ are disjoint four-column groups of the complement; each column of $W_7$ is their corresponding-column sum divided by $\sqrt{7}$. Add $W_b \operatorname{diag}(9, 8.5, 8, 7.5)W_b^\top$ to the common-$U$ covariance. The resulting local eigenvalues are $(9.4, 8.9, 8.4, 7.9)$, not the spike additions alone. For drift, $U_b = U\cos\alpha_b + V\sin\alpha_b$ with $\alpha_b = -36^\circ + 72^\circ b/7$. The influential family's block 7 instead uses $V$, eigenvalues $(100, 99, 98, 97)$ and $c = .01$. Exactly flat spectra and symmetric midpoint fixtures are deterministic abstention tests, not additional calibration families.

There are 256 calibration replicates per family. Orient each statistic by $w_j = s_j z_j$, with $s_j = 1$ for lower bounds and $-1$ for radii. Set

$$p_j = \min_{f \in \mathcal{F}_+} Q_{.10}(w_j \mid f), \qquad n_j = Q_{.90}(w_j \mid f_j^-), \qquad \theta_j = s_j(p_j + n_j)/2, \tag{14}$$

where $f_j^-$ is the sentinel in Table 8. $Q_p$ is the ascending order statistic $\lceil 256p \rceil$, with no interpolation: ranks 26 and 231 here. Every oriented gap must be at least .02. An abstention, incomplete record or numerical failure invalidates calibration. The joint rule must also accept at least 205/256 cases from

each valid family and at most 12/256 from each negative family. Only then are thresholds frozen and fresh validation opened.

Each of the ten validation families has 500 independent timelines. One-sided exact Clopper–Pearson bounds use $\alpha = .005$ each, a Bonferroni family-wise .05 allocation. Each valid lower bound must be at least .80 (at least 423/500 passes), and each invalid upper bound must be at most .05 (at most 12/500 false passes). Scientific abstentions stay in the denominator; unknown or numerical failures invalidate the run. All valid families had 500/500 passes and all invalid families 0/500. The bounds are .9894593118113242 and .010540688188675816. An earlier incomplete calibration attempt was retained as invalid; its numerical transient certificate was revised before this completed atlas, without changing the families, sample counts or selection rule.

The $K = 128$ bridge is another fixed-$r = 4$ test, with signed/permuted DCT-II orientations rather than Haar-QR. Its predeclared population lift cyclically extends complement spectra and scales active eigenvalues by the ratio of new to old complement trace (for isotropic complements, $124/28$). Local takeover scales its four common and four direct local eigenvalues by $(K - 8)/24$, retaining background .4. Near flat scales the five-dimensional boundary cluster by $(K - 5)/27$, retaining background 1.001 elsewhere. The changing-complement case uses its actual extended trace ratio. Thus this bridge is not evidence that unchanged thresholds work after arbitrary changes in dimension or signal-to-noise ratio.

### D.3 Adaptive rank selection: failure and sample-size follow-up

The later selector tries $r = 1, 2, 3, 4$ using shared thin SVDs and takes the largest passing rank, or zero. It first removes the global channel mean, $Y \leftarrow Y(I - \mathbf{1}\mathbf{1}^\top/K)$. Its D17/D18 synthetic paths are independent Gaussian coefficient tapes in a $K \times (K-1)$ orthonormal basis obtained by projecting an iid Gaussian matrix off $\mathbf{1}$ and applying thin QR. Paths are temporally centered and integrated, not sampled from the raw-timeline atlas. Both studies fix $K = 32, H = 16, B = 8$. Stable rank-$r$ families use the first $r$ SDs from $(3, 2.7, 2.4, 2.1)$ and .35 elsewhere. Six negatives use: all SDs 1; first four 1.2 and the rest 1; first five 2.5 and the rest .35; one SD 3 on axis $b$ in block $b$ and .35 elsewhere; block 0 with SDs $(3, 2.7, 2.4, 2.1)$ on the first four basis axes and .35 elsewhere, with all SDs 1 in the other seven blocks; or alternating rank-one and rank-four spikes in disjoint orthogonal basis directions. The latter uses SD 3 on axis 0 in even blocks and $(3, 2.7, 2.4, 2.1)$ on axes 4–7 in odd blocks, with .35 elsewhere.

D17 and D18 confirmation each use ten co-primary one-sided exact Clopper–Pearson bounds at $\alpha = .005$. Every stable family must recover its true rank in at least 175/200 replicates, with lower bound at least .80; each negative may select nonzero rank in at most 2/200, with upper bound at most .05. All ten conditions must pass. D17 used four paths per block and 200 replicates per family. It recovered ranks one through four in 200, 200, 199 and 169 cases, respectively. The rank-four lower bound .76847 failed the .80 criterion; all six negative families had zero false passes. D18 then planned block sample sizes $\{4, 6, 8, 12\}$ using, for each size, 100 independent rank-four replicates and 100 of each negative family. It selected the smallest size with at least 95/100 rank-four recoveries and no false passes in any negative family. Six paths gave 98/100. With that choice fixed, fresh 200-replicate families gave 200, 200, 200 and 197 correct rank recoveries, and zero false passes. The rank-four lower bound was .94620 and each negative upper bound .02614. These are distribution-specific operating results, not a universal six-path sufficiency claim. In the historical EEG application, the frame uses fit subjects S001–S032 only, eight fixed starts, $H = 16, K = 64$ and the recorded whole-subject RMS normalization. The later prefix-only sensitivity has its own preprocessing and does not inherit a fresh qualification claim.

### D.4 Historical generator gates

**Original retrospective EEG result, retained without revision.** The whole-record RMS normalization uses future increments; it is not a past-only evaluation. Let $S_{sa}$ be subject $s$'s historical off-diagonal Energy after averaging eight origins, and $R_{sa} = 1 - S_{s,\mathrm{cand}}/S_{sa}$. The 20,000 paired subject-bootstrap draws target $61^{-1}\sum_s R_{sa}$. Equal-subject improvements are 6.155% [4.208, 8.355] against the anchor, 9.121% [6.999, 11.466] against one-step TimeGrad, and 5.871% [3.900, 8.093] against Diffusion-TS-local. Pooled reductions are 6.084%, 9.190% and 5.800%, with 51, 58 and 49 of 61 subject wins. The candidate's anomaly residual is $9.337 \times 10^{-7}$, coverage er-

ror 0.02309 and normalized nearest-training distance 0.167712 against 0.05. Half-mix has slightly lower marginal CRPS, but coverage error 0.05828. These historical intervals condition on fixed fits and tapes, not training or model-selection uncertainty. The original Figure 4 is retained to distinguish this record from the newer main-text panel.

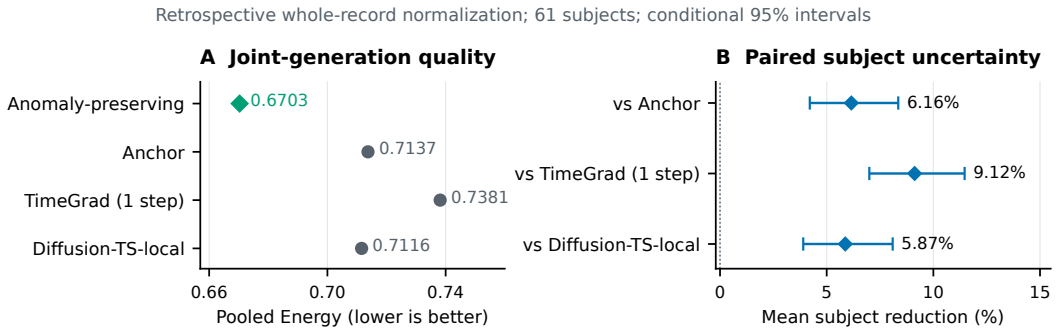


Figure 4: Original retrospective EEG comparison, not the prefix-only panel. Pooled off-diagonal Energy and mean subject-relative reductions have different estimands. The practical 1% gate applies to the pooled reduction.

The historical confirmatory statement is an intersection–union claim: the anomaly-preserving candidate must be better than every declared comparator. Each component passes the one-sided sign threshold and has a paired subject-bootstrap interval above zero; no multiplicity correction is needed for this single all-comparators conjunction. The separate 1% pooled-reduction rule is a practical gate. Individual pairwise results are not claimed as simultaneous field-wide superiority.

Table 10: Complete frozen-confirmation promotion registry (D25 experiment identifier). A missing value or failed row rejects the panel decision; a later row cannot rescue it.

| Gate | Frozen rule | Frozen result | Authorized interpretation |
|---|---|---|---|
| Source | S001–S032 only; methods and tapes fixed before confirmation | No refit | Proceed to eligibility |
| Eligibility | 61/61 finite and scored; anomaly error $\leq 5 \times 10^{-6}$ | 61/61; $9.337 \times 10^{-7}$ | Non-ranking screen |
| Relative | Pooled gain $\geq 1\%$; wins $\geq 39$; sign $p \leq 0.025$; bootstrap lower reduction $> 0$ against each | Minimum gain 5.800%; minimum wins 49; all intervals positive | Better than each declared comparator |
| No harm | Three secondary scores $\leq 1.01$ times the best comparator | All three lead | Safeguard passed |
| Calibration | Error $\leq$ best comparator $+0.01$ and $\leq$ half-mix; Energy within 1% of half-mix | 0.02309; Energy 0.192% better | Calibration passed |
| Novelty | Exact-copy rate 0; normalized distance $\geq 0.05$ | 0; 0.167712 | Narrow non-copy evidence |
| Control | Residuals $\leq 5 \times 10^{-11}$; identity exact; no dense action | Maximum $1.066 \times 10^{-14}$ | Exact geometric action |

Table 11: Complete-law safeguards and calibration. The half-mix is a mechanism ablation, not a primary comparator. Lower is better.

| Method | DCT CRPS | Variogram | Marginal CRPS | Coverage error |
|---|---|---|---|---|
| Synchronized anchor | 0.534798 | 0.120958 | 0.558570 | 0.03464 |
| TimeGrad | 0.553666 | 0.129114 | 0.579834 | 0.26087 |
| Diffusion-TS-local | 0.533215 | 0.120368 | 0.556491 | 0.04178 |
| Candidate half-mix ablation (D24) | 0.509699 | 0.116924 | **0.521620** | 0.05828 |
| Anomaly-preserving candidate (D24) | **0.508513** | **0.116156** | 0.521740 | **0.02309** |

**Whole-path non-copying.** For matching start $b$, let $\mathcal{T}_b$ be the S001–S032 training-path library and define $d(z, \mathcal{T}_b) = \min_{t \in \mathcal{T}_b} \|z - t\|_2 / \sqrt{HK}$. The registered novelty statistic is

$$N_{\mathrm{NN}} = \frac{\mathrm{median}_{b, z \in \mathcal{G}_b} \, d(z, \mathcal{T}_b)}{\mathrm{median}_{b, y \in \mathcal{C}_b} \, d(y, \mathcal{T}_b)}, \tag{15}$$

where $\mathcal{G}_b$ contains generated confirmation paths and $\mathcal{C}_b$ contains real S049–S109 confirmation paths. The observed value is 0.167712 against the frozen 0.05 threshold. No generated nearest-neighbour distance is exactly zero. This is a narrow whole-path non-copy test, not privacy evidence or a guarantee against copied subsequences.

## E SELECTIVE-RISK PROTOCOL AND DECISION SEMANTICS

The known-truth request study fixes $K \in \{32, 128\}$, $r \in \{1, 2, 4\}$, $H = 16$, and $M = 32$. Each $(K, r)$ cell contains 300 independent replicates in each of three truth classes. A replicate produces one frame/request decision and is the statistical unit; paths, members, times, channels, ranks, and certificate coordinates inside that replicate are repeated measurements. Safe nonidentity requests are assigned equally to scalar SPD gains in $\{0.85, 0.90, 1.10, 1.15\}$, rotated anisotropic SPD gains with eigenvalues in $[0.85, 1.15]$, and target-covariance requests. Contract-invalid replicates are assigned 50 each to an invalid frame, a low eigenvalue in $[0.65, 0.74]$, a high eigenvalue in $[1.26, 1.35]$, support distance in $[1.01, 1.25]$, action norm in $[0.21, 0.25]$, or an invalid covariance target. The remaining class contains 150 exact identities and 150 known-truth zero-rank frames per cell. Rank-selection misses and numerical failures remain in these fixed denominators.

The dimensionless support statistic $d$ is a typed policy input normalized so that the declared support boundary is one; the known-truth study sets safe $d = 0.4$ by construction and does not estimate a real-data support model. For base increments $X$ and proposed correction $\Delta$, the action statistic is

$$a = \frac{\|\Delta\|_F}{\max\{\|X\|_F, \epsilon_{\mathrm{tiny}}\}}, \tag{16}$$

where $\epsilon_{\mathrm{tiny}}$ is the smallest positive normal Float64 value. The production envelope requires $d \le 1$, $a \le 0.20$, and $0.75 \le \lambda_{\min}(G) \le \lambda_{\max}(G) \le 1.25$. A target covariance $T$ must be positive definite, satisfy $\lambda_{\min}(T) / \operatorname{mean}_j \lambda_j(T) \ge 10^{-4}$, and have condition number at most $10^4$. For $C_n = P_H X_n U$, the implementation uses $\Sigma = (NH)^{-1} \sum_{n=1}^{N} C_n^\top C_n$: centering is within each path over time, and the divisor is $NH$. The canonical Gaussian optimal-transport compiler (Takatsu, 2011) is

$$G = \Sigma^{-1/2} \big( \Sigma^{1/2} T \Sigma^{1/2} \big)^{1/2} \Sigma^{-1/2}, \tag{17}$$

followed by the same gain, support, action, and post-action certificate gates.

This expression is not evaluated through a pseudoinverse. For $s = \max\{\|\Sigma\|_2, \|T\|_2, \epsilon_{\mathrm{tiny}}\}$, the numerical compiler requires strictly positive eigenvalues above $10^{-12}$ for both $\Sigma/s$ and $(\Sigma/s)^{1/2}(T/s)(\Sigma/s)^{1/2}$. Finite, symmetric inputs are required (symmetry tolerance $10^{-12}$). A singular, indefinite or numerically degenerate source covariance therefore causes ABSTAIN and an unchanged path copy; it is not regularized, silently down-ranked or repaired. Direct gain requests do not require covariance inversion. Schema or nonfinite caller errors remain errors, not successful scientific decisions.

The descriptive risk–coverage sweep fixes $\tau \in \{-0.20, -0.10, -0.05, 0, 0.05, 0.10, 0.20\}$ and never selects a new production threshold. At each $\tau$, safe coverage is safe applies divided by the same 1,800 safe replicates, false-apply rate is invalid applies divided by the same 1,800 invalid replicates, and selective risk under the registered equal-mass mixture is invalid applies divided by all safe-plus-invalid applies. At $\tau = 0$, removing any one frame gate still produces zero false applies because the raw frame negatives violate overlapping gates; removing one request gate admits 295–300 invalid replicates. These are redundancy diagnostics, not claims that a gate uniquely captures a physical failure.

The official seed 2708241 was executed once, without retry, replacement, or threshold change. The 5,400 replicates contain no real-data read and zero model training updates. The run used eight local arm64 CPU processes for 43.066 seconds, without CUDA, Metal, or cloud GPU. Separate

deterministic core tests verify bit-exact identity and the thin composition and inverse identities to absolute tolerance $2 \times 10^{-13}$; they do not add statistical units, and the production selective policy remains restricted to SPD gains in its frozen envelope.

## F DATA-USE CHRONOLOGY AND IMPLEMENTATION

EEGMMIDB was collected using BCI2000 (Schalk et al., 2004) and is distributed by PhysioNet (Pollard et al., 2026) under the dataset provider's Open Data Commons Attribution License v1.0. We use its R02 recordings only; this attribution does not extend the evaluated data scope.

Table 12: EEGMMIDB-R02 chronology. The subject is the inferential outer unit.

| Phase | Subjects | Permitted use | Boundary |
|---|---|---|---|
| Fit | S001–S032 | Fit the candidate, TimeGrad, Diffusion-TS-local, and the frame | Checkpoints, samplers, and seeds stored |
| Trust | S033–S040 | One fixed candidate pass/fail gate | No tuning grid |
| Development | S041–S048 | One pass/fail opening qualification | No confirmation refit or threshold change |
| Confirmation | S049–S109 | Score all 61 subjects once | Methods, tapes, endpoints, and thresholds fixed |

Each R02 subject contributes starts at 2, 8, 14, 20, 26, 32, 38, and 44 seconds and the same 16 target increments. The candidate uses 32 context increments; TimeGrad and Diffusion-TS-local use 96-increment histories. One whole-run increment-RMS normalization is applied per subject; channels are not selected, filtered, resampled, or interpolated. Specifically, for the supplied $T$-sample recording $x$, the scalar is

$$s(x) = \left[ \frac{1}{(T-1)K} \sum_{t=1}^{T-1} \sum_{k=1}^{K} (x_{t+1,k} - x_{t,k})^2 \right]^{1/2}, \qquad \widetilde{x} = x/s(x).$$

The implementation computes this scalar before extracting history and target windows. Changing only a future suffix can therefore change the normalized history. Holding out subjects does not make this preprocessing past-measurable: the reported result concerns retrospective whole-record-normalized paths, not a deployable online forecast with a training-only scale. We preserve the registered scores and decisions. Appendix F.1 reports a separate post-hoc prefix-only matched-retraining analysis, not a retroactive correction or new confirmation of the historical experiment. The ensemble-dependent common shift also couples members, so the retained off-diagonal score does not by itself establish the population-risk sampling assumptions in Equation 1. The separate prefix-normalized and finite-measure analyses do not retroactively replace this result. The EEG TimeGrad implementation uses the pinned ProbTS numerical core (Zhang et al., 2024), with 1,000 AdamW updates and only one diffusion step per autoregressive forecast time, using a constant $\beta = 0.99438123898062625$. Its four-layer, width-128 GRU uses context 96; lags, time features, feature-index embeddings, and internal scaling are disabled. At $H = 16$, the sampler makes 16 denoiser calls per path. Matching the candidate's 16 field evaluations does not match FLOPs, wall time, or optimization adequacy. Here and in the prefix analysis, the table label TimeGrad denotes this budget-limited configuration, not a reproduction of the original paper's best-performing diffusion schedule. A separately labeled 100-step sensitivity is reported below. Diffusion-TS-local is an $\epsilon$-prediction model with trend/season branches in the rank-1 coordinate, 1,000 AdamW updates, and a 16-step deterministic DDIM sampler (Song et al., 2021). It retains the 63-dimensional complement and is not an official dense Diffusion-TS reproduction. In particular, its noise-prediction loss does not implement the original model's sample reconstruction and Fourier-loss objective (Yuan & Qiao, 2024). Equal training-update counts do not establish equal convergence or equivalence to either published method's strongest configuration.

The proposal is learned by conditional flow matching (Lipman et al., 2023). For target path $Z_1$, a synchronized path $Z_0$ from a different fit subject at the same start, history $c$, and $t \sim U[0,1]$, let

$Z_t = (1-t)Z_0 + tZ_1$ and minimize

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\left[\frac{1}{HK}\left\|v_\theta(Z_t, t, c) - (Z_1 - Z_0)\right\|_F^2\right]. \tag{18}$$

The terminal proposal $F$ is obtained by integrating the learned field from a synchronized anchor with eight fixed Heun steps. The model uses context 32, horizon 16, width 64, 16 inducing tokens, four attention heads, two blocks, and a zero-initialized output layer. AdamW runs 1,000 updates with batch size 16, learning rate $2 \times 10^{-4}$, weight decay $10^{-4}$, clipping at 1.0, and seed 240823. Sampling uses eight Heun steps (16 field evaluations) and $M = 32$. Each reported configuration uses one fit per neural method: seed 240823 for the candidate, 250823 for TimeGrad, and 250824 for Diffusion-TS-local. Seed 250825 generates fixed per-model forecast tapes, and seed 250826 generates the subject-bootstrap resamples; neither is a training fit.

### F.1 Post-hoc prefix-calibration sensitivity

The historical whole-record RMS uses future increments. We therefore retrained the disclosed budget-limited comparators with past-only calibration. This post-hoc sensitivity does not replace D25 or make its already-used subjects a fresh confirmation set. For each recording we freeze $s = \sqrt{(320K)^{-1}\sum_{j=1}^{320}\sum_{k=1}^{K}(X_{j,k} - X_{j-1,k})^2}$ using only levels 0 through 320, available at the first origin. Each history and target increment is divided by this same scalar; the first future level is 321. The information set contains the recent history *and* this two-second calibration prefix. No future-based clipping, imputation, subject removal, or calibration-length search is performed.

All FIT windows, the rank-one basis, anchor library, and coordinate statistics were rebuilt from S001–S032 with this rule. Flow, TimeGrad, and Diffusion-TS-local were each retrained from initialization for exactly 1,000 AdamW updates with batch 16 and the same respective training seeds, optimizer settings, and sampler budgets described above; no HPO, early stopping, or checkpoint selection was added. Flow uses context 32; TimeGrad and Diffusion-TS-local use context 96. The two candidate variants reuse the same flow proposal. Rank remains fixed at one, without repeating rank selection or claiming that the old rank-qualification gate transfers. S033–S048 were not reused for selection. We retain the 61 S049–S109 subjects, eight origins, $H = 16$, $K = 64$, and $M = 32$. Independent baseline noise tapes use the original call layouts; equal seeds are not claimed to provide cross-model common random numbers. The additional 3,000 training updates ran on shared local MPS with a frozen ordered CPU reduction for the diffusion model's small regime embedding, without changing its forward lookup or optimizer. Sharing does not establish unaffected performance.

Table 13: Post-hoc R02 prefix-calibration sensitivity, not fresh confirmation. TimeGrad uses one denoising step per forecast time; Diffusion-TS-local is a rank-one adaptation, not the official dense model. Eight origins are averaged within each of 61 equally weighted subjects. $E_{\mathrm{off}}$ retains the off-diagonal pair denominator; $E_{M^2}$ scores the finite empirical ensemble with the $M^2$ denominator. Both are lower better; neither coupled-member score is presented as an unbiased population risk estimator. Coverage error is averaged at the origin level.

| | $E_{\mathrm{off}}$ | | $E_{M^2}$ | | Coverage |
|---|---|---|---|---|---|
| Method | Prefix | $\mu$V | Prefix | $\mu$V | error |
| Synchronized anchor | 0.680871 | 17.099944 | 0.701926 | 17.625447 | 0.074214 |
| TimeGrad | 0.698257 | 17.571955 | 0.711154 | 17.894823 | 0.248227 |
| Diffusion-TS-local | 0.678704 | 17.046133 | 0.699504 | 17.564325 | 0.075576 |
| Half-mix ablation | 0.639528 | 15.429263 | 0.658202 | 15.895806 | 0.084633 |
| Anomaly-preserving | 0.638868 | 15.450631 | 0.659923 | 15.976134 | 0.070117 |

The primary estimand is prefix-normalized $E_{\mathrm{off}}$. Candidate reductions in the ratio of mean scores are 6.169% [4.187, 8.423] against the anchor, 8.505% [6.255, 11.032] against TimeGrad, and 5.869% [3.862, 8.165] against Diffusion-TS-local, with 51, 57, and 51 subject wins out of 61. Physical-unit reductions are 9.645% [4.719, 15.258], 12.072% [6.674, 18.090], and 9.360% [4.419, 14.936], respectively. Intervals use 20,000 paired subject-bootstrap draws with seed 250826 and shared indices across comparisons. They are descriptive marginal percentile intervals, not simultaneous bounds or new confirmatory tests. The ratio of means here differs from D25's mean of subject-wise ratios;

subtracting these percentages does not estimate the causal effect of normalization. Fits, tapes, and the reused evaluation cohort remain fixed.

The gains over the three primary comparators persist under both score denominators and both units. The candidate does not dominate the half-mix ablation. Half-mix has lower physical-unit $E_{\mathrm{off}}$ and lower $E_{M^2}$ in both units, as well as lower prefix marginal CRPS (0.497309 versus 0.497837). The candidate has lower origin-level coverage error (0.070117 versus 0.084633). Its prefix DCT-CRPS and variogram scores are 0.483853 and 0.110489, compared with half-mix's 0.484791 and 0.111155. Thus the quality–calibration trade-off persists; the main endpoint was not changed after observing these secondary comparisons.

All 488 origins pass the original anomaly and three-request control certificates; the maximum anomaly error is $1.848 \times 10^{-6}$ against the unchanged $5 \times 10^{-6}$ tolerance. Control probes are not newly scored arms. The candidate's exact whole-path copy rate is zero, whereas the synchronized anchor has rate one by construction. The ratio of mean origin-level median nearest-training distances is 0.182631 in prefix units and 0.217723 in physical units. These origin-level coverage and novelty reductions differ from the historical pooled definitions, so we do not reapply or claim a new pass of the D25 conjunction.

A read-only terminal audit verified all 488 sample/score commits, 61 batch addresses, fixed source and input bindings, finite metrics, dual-unit consistency, and the certificates from saved samples. Latest model/tape counters and all active AdamW step counters equal 1,000; this is not an independent per-update historical journal. Subject aggregation and the original bootstrap were independently recomputed from saved scores and matched the summary. The audit reused scientific certificate validators and did not independently rescore raw EDF targets. This study supports robustness to the specified past-only RMS preprocessing, not field-wide SOTA, new-cohort confirmation, training-seed robustness, or causality of the acquisition hardware's preprocessing. No additional EEG run or sealed GPM year was opened.

### F.2 Post-hoc TimeGrad diffusion-schedule sensitivity

The main text now displays the finite-ensemble score from this complete panel. All original off-diagonal results below remain unchanged. The additional inference at the end of this subsection is a labelled post-hoc analysis of saved subject averages, not a new experiment or promotion decision.

A separate matched-budget retraining tests how the TimeGrad diffusion schedule affects the EEG generation comparison. We retain the prefix-only FIT, 61 evaluation subjects, eight origins each, context 96, $H = 16$, $K = 64$, $M = 32$, GRU architecture, seed 250823, and 1,000 AdamW updates with batch 16. The new schedule uses 100 linearly spaced betas from $10^{-4}$ to $0.1$. The old constant beta matches its terminal forward-noise variance, but not its intermediate training-noise levels or reverse chain. This is a schedule intervention, not an assertion that the two models are equivalent. All trainable parameters start identically. The original training-window indices, Gaussian training noise, and forecast initial noise are reused. New training diffusion indices and 99 reverse innovations per forecast time are independently addressed and persisted before use; the final reverse step adds no innovation. The new sampler executes 1,600 denoiser calls per path rather than 16. Equal update budgets do not establish equal convergence, FLOPs, or sampling cost.

Only the new TimeGrad model is trained. The other five methods and their scores are reused unchanged from Appendix F.1. No additional HPO, early stopping, checkpoint selection, seed selection, or evaluation cohort is introduced. The 488 new sample/score commits are complete before this analysis. This same-cohort, single-initialization experiment is post-hoc schedule sensitivity, not fresh confirmation or a reproduction of the original TimeGrad paper's strongest configuration.

The multistep schedule improves TimeGrad's prefix $E_{\mathrm{off}}$ from 0.698257 to 0.677704, a 2.943% reduction in the ratio of means. The candidate remains lower at 0.638868, a 5.731% reduction relative to TimeGrad (100), compared with 8.505% against TimeGrad (1). Physical-unit reductions are 2.438% for TimeGrad (100) relative to (1), and 9.875% for the candidate relative to TimeGrad (100). These percentages are point ratios, not the endpoints of the paired-difference intervals in Table 16. Thus the one-step schedule materially affects the comparison, while the candidate's advantage persists in this specified matched-budget analysis. It does not establish superiority over a fully optimized TimeGrad.

Table 14: Complete six-method score panel after the TimeGrad schedule sensitivity. TimeGrad (1) and (100) denote denoising steps per forecast time. All scores use the same frozen targets and 61 equally weighted subject averages; lower is better. Cov. is mean origin-level coverage error. Physical-unit rows retain dimensionless coverage error.

| Method | $E_{\mathrm{off}}$ | $E_{M^2}$ | DCT-CRPS | Variogram | Marg. CRPS | Cov. |
|---|---|---|---|---|---|---|
| *Prefix-normalized* | | | | | | |
| Anchor | 0.680871 | 0.701926 | 0.509407 | 0.115163 | 0.533390 | 0.074214 |
| TimeGrad (1) | 0.698257 | 0.711154 | 0.524584 | 0.123078 | 0.549155 | 0.248227 |
| Diffusion-TS-local | 0.678704 | 0.699504 | 0.507909 | 0.114641 | 0.531310 | 0.075576 |
| Half-mix | 0.639528 | 0.658202 | 0.484791 | 0.111155 | 0.497309 | 0.084633 |
| Anomaly-preserving | 0.638868 | 0.659923 | 0.483853 | 0.110489 | 0.497837 | 0.070117 |
| TimeGrad (100) | 0.677704 | 0.693891 | 0.508877 | 0.121095 | 0.529521 | 0.131917 |
| *Physical units ($\mu$V)* | | | | | | |
| Anchor | 17.099944 | 17.625447 | 12.445833 | 2.806359 | 13.520083 | 0.074214 |
| TimeGrad (1) | 17.571955 | 17.894823 | 12.861719 | 3.038269 | 13.961653 | 0.248227 |
| Diffusion-TS-local | 17.046133 | 17.564325 | 12.406179 | 2.798799 | 13.450708 | 0.075576 |
| Half-mix | 15.429263 | 15.895806 | 11.584401 | 2.695066 | 12.019084 | 0.084633 |
| Anomaly-preserving | 15.450631 | 15.976134 | 11.585967 | 2.697435 | 12.072913 | 0.070117 |
| TimeGrad (100) | 17.143522 | 17.548708 | 12.512768 | 2.977028 | 13.517669 | 0.131917 |

Table 15: Remaining saved diagnostics for all six methods. Distances are means of origin-level median nearest-training RMS distances, not diversity objectives with a preferred monotone direction. Exact copy is a whole-path training-reference copy rate; the anchor copies by construction.

| Method | Generated distance | Truth distance | Exact copy |
|---|---|---|---|
| *Prefix-normalized* | | | |
| Anchor | 0.000000 | 1.025823 | 1.000000 |
| TimeGrad (1) | 0.867443 | 1.025823 | 0.000000 |
| Diffusion-TS-local | 0.552498 | 1.025823 | 0.000000 |
| Half-mix | 0.237592 | 1.025823 | 0.000000 |
| Anomaly-preserving | 0.187347 | 1.025823 | 0.000000 |
| TimeGrad (100) | 0.946197 | 1.025823 | 0.000000 |
| *Physical units ($\mu$V)* | | | |
| Anchor | 0.000000 | 24.660290 | 1.000000 |
| TimeGrad (1) | 21.679047 | 24.660290 | 0.000000 |
| Diffusion-TS-local | 13.735565 | 24.660290 | 0.000000 |
| Half-mix | 6.548290 | 24.660290 | 0.000000 |
| Anomaly-preserving | 5.369107 | 24.660290 | 0.000000 |
| TimeGrad (100) | 23.620863 | 24.660290 | 0.000000 |

The complete six-method, nine-metric panel retains the candidate's non-dominance over half-mix. Half-mix remains lower on physical $E_{\mathrm{off}}$, $E_{M^2}$ in both units, and marginal CRPS in both units, and physical DCT-CRPS and variogram. The candidate has lower coverage error. TimeGrad (100) reduces coverage error to 0.131917 but remains worse on that diagnostic than the other four non-TimeGrad methods. It has lower prefix $E_{\mathrm{off}}$ than Diffusion-TS-local but higher physical-unit $E_{\mathrm{off}}$, so even these baseline rankings depend on the reporting unit. The candidate's generated/truth nearest-training-distance ratios remain 0.182631 (prefix) and 0.217723 (physical), versus 0.922379 and 0.957850 for TimeGrad (100). Zero exact copies do not eliminate proximity to the training library. No endpoint or model is selected anew using these secondary results.

The independent read-only terminal audit verified all 122 forecast-batch bindings and 488 contiguous origins, 92 files in the new source binding and 87 original frozen source files, unchanged FIT and reused tapes, finite metrics, dual-unit consistency, 32 distinct new members and zero exact reference copies at every origin. The original candidate control certificates remain valid and unchanged; they are not attributed to the new TimeGrad baseline. The checkpoint cursor, all active AdamW step counters, and loss-history length are 1,000. These checks are consistent with the sequential trainer but are not an independent historical trace of every update. Independently im-

Table 16: The two predeclared $E_{\text{off}}$ contrasts, in both reporting units. Negative differences favor the left method. Intervals are descriptive marginal 95% percentile intervals from 20,000 paired subject-bootstrap draws (seed 250826), with the same draw indices across contrasts. Eight origins are averaged within each person before resampling 61 people; these are not simultaneous bounds or confirmatory tests.

| Contrast | Unit | Mean difference | 95% interval |
|---|---|---|---|
| TimeGrad (100) − TimeGrad (1) | Prefix | -0.020553 | [-0.023891, -0.017121] |
| TimeGrad (100) − TimeGrad (1) | $\mu$V | -0.428433 | [-0.546655, -0.293041] |
| Candidate − TimeGrad (100) | Prefix | -0.038836 | [-0.058101, -0.022261] |
| Candidate − TimeGrad (100) | $\mu$V | -1.692892 | [-3.432990, -0.565074] |

plemented 61-subject aggregation and the original bootstrap match the saved summary. The audit reuses score/certificate validators and does not rescore raw EDF targets or independently recompute scores from sample/target pairs. Stored tape addresses, layout and bound generation code are checked, but random values are not independently regenerated to prove provenance or consumption. No EEG R03–R14 or sealed GPM year is opened. Neither this audit nor the favorable contrast establishes training-seed robustness, a new confirmation gate, or field-wide SOTA.

#### F.2.1 Issued-ensemble Energy: paired inference from saved aggregates

For this revision we fixed a bounded post-hoc analysis before executing it: use only the already audited $61 \times 2 \times 6 \times 9$ subject-score tensor, compare the candidate with all five alternatives in both original units, and retain the original 20,000 paired subject-bootstrap addresses (seed 250826). We read no raw targets or sample paths and perform no fitting, sampling, rescoring or model selection. The eight origins were already averaged within each subject. For each comparator $a$, the statistic is $61^{-1}\sum_s(V_{s,\text{cand}} - V_{sa})$, where $V$ is exact Energy of the issued empirical ensemble, with denominator $2M^2$. Each draw resamples 61 subjects with replacement. An independently count-weighted calculation agrees with the index-based calculation within $2 \times 10^{-13}$. Input bytes, terminal status and aggregate means remain unchanged. This verifies the added arithmetic, not the original raw-data scoring.

Table 17: Complete post-hoc V-score contrasts. Negative differences favor the anomaly-preserving candidate. Marginal 95% percentile intervals are conditional on the saved fits and tapes; they are not simultaneous bounds. Reduction is a point ratio of equally weighted subject means. Wins count subjects, not individual origins or paths.

| Comparator | Unit | Cand. minus ref. | 95% interval | Reduction, % | Wins/61 |
|---|---|---|---|---|---|
| Anchor | Prefix | -0.042003 | [-0.057323, -0.028522] | 5.984 | 51 |
| TimeGrad (1) | Prefix | -0.051231 | [-0.069338, -0.035215] | 7.204 | 53 |
| Diffusion-TS-local | Prefix | -0.039581 | [-0.055086, -0.025987] | 5.658 | 51 |
| Half-mix | Prefix | 0.001721 | [0.000302, 0.003193] | -0.261 | 23 |
| TimeGrad (100) | Prefix | -0.033968 | [-0.053272, -0.017358] | 4.895 | 44 |
| Anchor | $\mu$V | -1.649313 | [-3.062565, -0.720190] | 9.358 | 51 |
| TimeGrad (1) | $\mu$V | -1.918689 | [-3.523895, -0.853822] | 10.722 | 53 |
| Diffusion-TS-local | $\mu$V | -1.588191 | [-2.976519, -0.664822] | 9.042 | 51 |
| Half-mix | $\mu$V | 0.080328 | [0.020752, 0.159516] | -0.505 | 23 |
| TimeGrad (100) | $\mu$V | -1.572574 | [-3.293532, -0.457118] | 8.961 | 44 |

The candidate's favorable contrasts against the three neural configurations and anchor persist for the issued-ensemble score. Half-mix is lower in both units, and its marginal difference intervals exclude zero in the opposite direction. Thus the revised score object does not support candidate dominance over all six methods. The candidate retains the lower coverage error already reported in Table 14. The old off-diagonal results and promotion decision are not overwritten. These intervals describe subject variation for one saved experiment; they exclude training-seed, tuning, population-shift and model-selection uncertainty.

## G COMPLETE RESULT TABLES AND NEGATIVE STUDIES

**Credit correction only for what it adds.** Four support-native backbones were fitted on 2018 and adapted on 2019 before 2020 development. Each Q2 preserves its Q1 ensemble mean, adjacent-pair differences and support floor, not KRPC's endpoint invariant. Three adapter initializations share 256 training origins, 300 Q1 and 400 Q2 updates. Origins, not adapters or paths, are the inferential units. The development summary (Table 21, Appendix G) separates proposal, calibration and correction. TimeGrad Q1 improves Energy by 27.45% over its earlier shift-only TISS; Q2 adds smaller, backbone-dependent gains. D3U's Q2 interval crosses zero. TimeGrad and TACTiS-2 worsen increment Energy by 2.17% and 4.11%; TACTiS-2 worsens q99 error by 8.33%. Appendix L gives complete profiles, including positive D3U increment/variogram results.

**An explicit primary-objective trade-off.** On exposed 2020 development, tail-aware training lowers path-q99 CRPS by 0.3145% versus an equal-update Energy-only policy at a 0.0030% Energy cost. The tail gain has the same sign in three technical replicates and twelve months. It ranks third of thirteen on that score, behind tail specialists: an objective-sensitive trade-off, not new confirmation or SOTA (Appendix O.4).

**A real-data mechanism check, not another quality claim.** On ten GHCNh station pairs, a separately frozen 2024 relation-weight rule retains weight at least 0.5 in 82/83 natural windows. Primary reversal, coordinate mismatch and time mismatch each receive weight at most 0.05 in 83/83 windows. Shared distortions remain invisible. This supports the observability mechanism at $K = 10$, without evaluating a future or proving quality improvement (Appendix P).

**Distinguishing legitimate change from structural failure.** The initial challenge qualifies named phase, lag and copying failures (Appendix X). A new comparison uses independent draws, eight statistics and family-specific nominal 1% levels. Phase–lag moments average 80.396% detection across eight fixed dynamic conditions, with largest legal false-positive upper bound 3.423%. They meet the prespecified five-percentage-point advantage over both raw-path Energy references and two variogram references, but not phase-pooled variogram (79.517%) or Energy on the same dynamic features (84.595%). Conversely, full-path Energy detects every common-shift case; the centered statistics are blind to that change. This supports complementary, task-matched endpoints, not a uniquely best metric (Appendix Z).

The separate 5,400-case selective-risk panel covers $K \in \{32, 128\}$ and $r \in \{1, 2, 4\}$. Safe coverage is 0.9928; no invalid request applies in 1,800 trials, and all 1,800 identity/zero-rank requests return bit-exact NO-OP. All 18 Bonferroni assertions pass. Request-only, frame-only and reachability screening falsely apply at rates 0.1667, 0.6628 and 0.8333 (Figure 5). These results qualify the named synthetic truth classes, not every real-data request.

**Low-rank real-data execution.** On EEG, the selected rank-one action protects 63 directions, with endpoint and complement residuals below $5 \times 10^{-15}$.

**Joint success on cached real paths.** On 90 Solar validation ensembles ($M = 100, H = 24, K = 137$), a post-hoc task requests constant or phase-reversing relations between two public channel baskets, while protecting each phase's channel sums and basket complement. After a separately tested common-unit numerical repair, KRPC achieves 180/180 joint successes, tied with protected whiten–color and dense OT. Scalar/diagonal controls miss the requests; unrestricted affine control breaks the phase sums. The original implementation's 142 numerical refusals remain reported (Table 63). Secondary path changes are allowed costs, not hidden failures. This demonstrates specified control, not a new generator-quality lead or real grid benefit.

### G.1 WHAT DO THE MATCHED CONTROLS ESTABLISH?

Path conditioning adds value in the 2024 matched weather comparison, and request geometry adds value in the $K = 500$ operator study. Execution itself need not be unique: SVD projection reproduces graph protection, and dense OT ties KRPC on the real-cache task. These comparisons support a reusable interface rather than an irreplaceable transform.

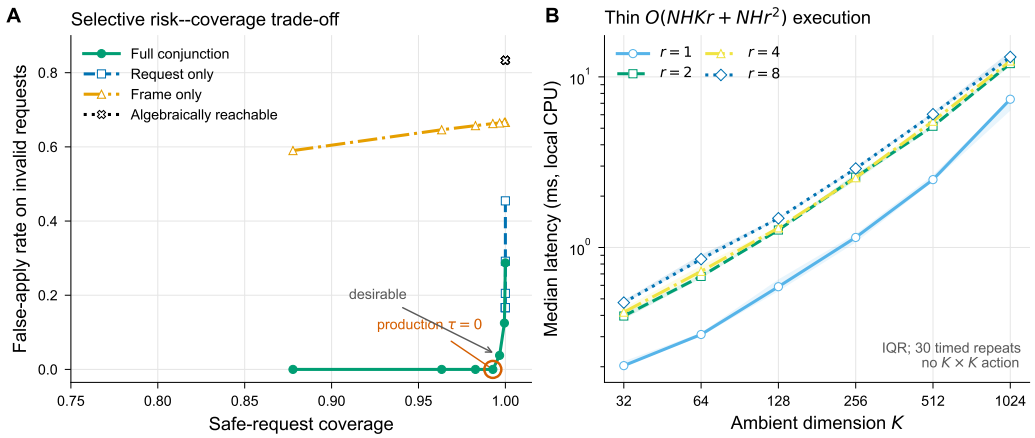


Figure 5: **Qualification tests both action and refusal.** The production conjunction trades 0.72 percentage points of safe coverage against request-only screening while reducing false applies from 16.67% to zero. Median latency and IQR use 30 local-CPU repeats; no dense $K \times K$ action is formed. The left panel tests the request policy, not generation quality; the right measures execution cost, not cross-hardware superiority.

Other domains delimit quality transfer. On GHCNh, the plugin improves its CSDI base by 3.38% Energy but stronger scalar/static heads improve more. Solar's marginal-preserving permutations trail fixed simple controls on joint and aggregate scores. OISST and PTB-XL likewise show that exact action or better point error need not improve the generated law. Complete comparisons and spread-collapse diagnostics remain in Appendix G and the domain-specific appendices.

### G.2 Domain-specific limits

**Evidence limits.** Certificates establish specified identities, not truth-law equality, causal benefit or privacy. The relation repair still misses common-mode distortion; adaptation to arbitrary nonstationarity remains unproved. EEG's original normalization uses future information; its prefix-only sensitivity reuses the cohort. EEG and USGS have one neural fit per method. USGS retains temporal dependence and 3.17% negative D3U values. Coupled U-scores have no established unbiased marginal-law interpretation. MERRA-2's support-native backbones and conditional block intervals have the qualifications in Appendices B and L. GHCNh lacks some targets; Solar has seven test days and longer control histories. The full attribution and unresolved or adverse results are retained in Appendices Q and P, including Solar residual and GPM selection studies.

Table 18: Primary EEGMMIDB-R02 comparison. Lower Energy is better.

| Comparator | Energy | Pooled reduction | Mean subject reduction [95% CI] | Wins |
|---|---|---|---|---|
| Synchronized anchor | 0.713730 | 6.084% | 6.155% [4.208, 8.355] | 51/61 |
| TimeGrad | 0.738144 | 9.190% | 9.121% [6.999, 11.466] | 58/61 |
| Diffusion-TS-local | 0.711576 | 5.800% | 5.871% [3.900, 8.093] | 49/61 |
| Anomaly-preserving candidate (D24) | 0.670306 | – | – | – |

The panel contains the first 500 stations with complete finite daily records in a preordered, HUC2-balanced 1,024-site roster. Station magnitudes do not enter selection. Its $9{,}497 \times 500$ daily-mean array is used without imputation, interpolation or clipping. The calendar split is fit 2000–2018, development 2019–2020, one-shot qualification 2021 and outer confirmation 2022–2025. Context is 96, horizon 16 and ensemble size $M = 64$. D3U leads development by 1.901% and the one-shot 2021 comparison by 3.870% against dynamic factor. Each origin's Energy uses flattened standardized paths, with both norm terms divided by $\sqrt{HK}$. The 50,000-replicate bootstrap resamples 48 calendar-month units, each containing eight overlapping origins. All eight comparisons pass the

Table 19: USGS $K = 500$ outer confirmation over 384 origins. Lower fair normalized Energy is better. The reduction column is relative to D3U. All eight paired comparisons pass the 0.5% practical gate and simultaneous Bonferroni-adjusted one-sided inference.

| Arm | Mean Energy | D3U reduction |
|---|---|---|
| D3U | 0.440502 | – |
| Dynamic factor | 0.457125 | 3.6366% |
| TACTiS-2 | 0.460919 | 4.4298% |
| Calendar block | 0.513506 | 14.2169% |
| CSDI | 0.525992 | 16.2531% |
| TimeGrad | 0.699255 | 37.0041% |
| Last value | 0.701884 | 37.2400% |
| Weekly persistence | 0.777068 | 43.3123% |
| TimePrism (Dai et al., 2026) | 1.473165 | 70.0983% |

predeclared 0.5% practical gate and a Bonferroni-adjusted one-sided lower bound above zero. This task-specific contract retains source, eligibility, grouping and effect gates. It does not rerun EEG's entire calibration–novelty–control conjunction on streamflow.

D3U ranks first in every outer year from 2022 through 2025; the closest Bonferroni-adjusted lower bound is 0.007460. Toto is retained in the execution record as target-free attempted-ineligible after a nonfinite physical inverse at the common $M = 64$ interface. It is not assigned an infinite score, repaired, or silently removed from the declared roster.

The unchanged claim-safety diagnostics also expose why raw-law realism remains closed: across outer samples, D3U reaches a physical maximum of $1.77 \times 10^{22}$ and minimum of $-5.14 \times 10^{5}$, with 3.17% negative values; TimePrism reaches $6.96 \times 10^{179}$. These diagnostics do not alter the predeclared standardized-space ranking, but they prohibit a physical-tail or deployment-readiness claim.

All fitted USGS arms use only the 6,940 rows from 2000–2018. D3U uses the released point-plus-residual construction with 2,140 point updates and 21,400 diffusion updates. TimeGrad and full-$K$ CSDI use 5,000 updates each; TACTiS-2 uses 2,500 marginal/flow and 2,500 copula updates; and TimePrism Short uses 6,000 updates with its released 625-scenario product construction. Dynamic factor selects rank from $\{8, 16, 32, 64\}$ on fit-internal 2016–2018 origins; the remaining classical arms have no development-selected hyperparameter. Every neural arm contributes one strictly reloaded terminal checkpoint, with no validation selection, early stopping, post-development refit, or budget extension. These are method-specific released/adapted contracts, not a compute-matched architecture tournament.

The requested arm is 0.089264 times the matched-unrequested mean. Identity is bit-exact and all coordinate, protected-complement and endpoint certificates pass below $2.22 \times 10^{-15}$. This table contains no realized-future or forecast-quality score.

Table 20: Selective-KRPC co-primary results. Bounds are one-sided exact Clopper–Pearson intervals after Bonferroni allocation over all 18 assertions.

| $K$ | $r$ | Safe apply | Lower | False apply | Upper | Correct no-op | Lower |
|---|---|---|---|---|---|---|---|
| 32 | 1 | 300/300 | 0.980571 | 0/300 | 0.019429 | 300/300 | 0.980571 |
| 32 | 2 | 300/300 | 0.980571 | 0/300 | 0.019429 | 300/300 | 0.980571 |
| 32 | 4 | 296/300 | 0.955991 | 0/300 | 0.019429 | 300/300 | 0.980571 |
| 128 | 1 | 300/300 | 0.980571 | 0/300 | 0.019429 | 300/300 | 0.980571 |
| 128 | 2 | 300/300 | 0.980571 | 0/300 | 0.019429 | 300/300 | 0.980571 |
| 128 | 4 | 291/300 | 0.931414 | 0/300 | 0.019429 | 300/300 | 0.980571 |

The PTB-XL sequence uses only folds 1–7 after the permanent unscored fold-8 abort. D31 improves the legacy synchronized anchor by 3.437% through a richer history-only temporal-phase signature, while its learned flow selects identity. D32, D33, and D34 improve the retained source by 0.422%, 0.598%, and 0.272% but fail the frozen 1% practical gate. D35 opens empirical support with a free whitened residual and worsens Energy by 8.973%. D36 reranks by a predicted future-response point

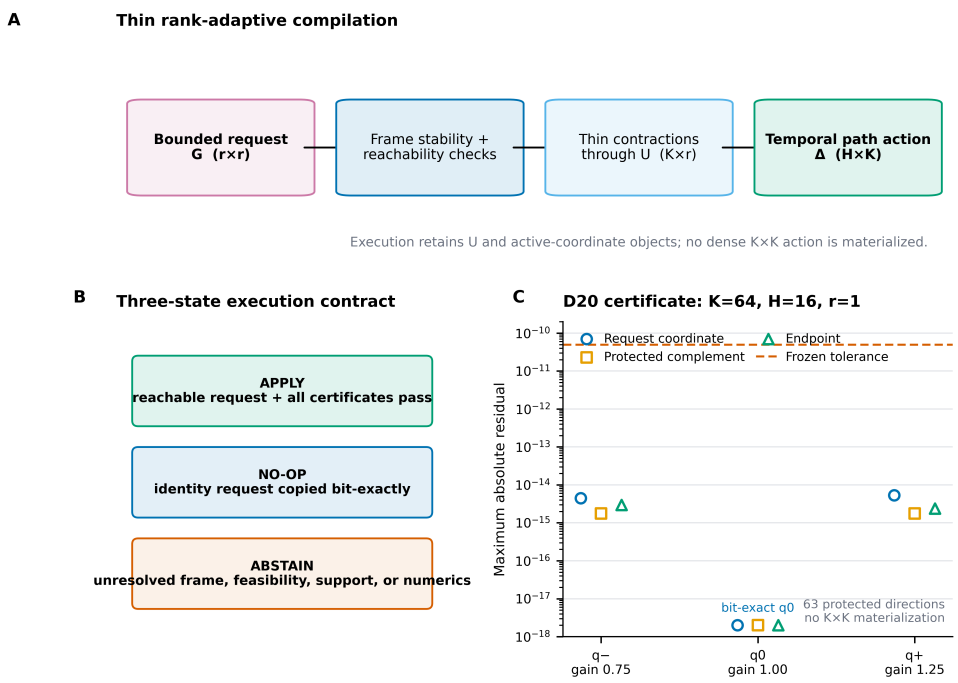


Figure 6: Thin rank-adaptive control and deterministic residual certificates at $K = 64, H = 16, r = 1$. Exact zero residuals are placed below the positive log floor for visibility only.

Table 21: **Separate the proposal, calibration and protected correction.** Matched MERRA-2 development comparison. Lower Energy V is better. Q2 gain is relative to its own Q1; brackets give the original conditional intervals (99.642857% for CSDI/TimeGrad, 99.583333% for D3U, 99.666667% for TACTiS-2; separate comparison families). NativeRootAR Energy is 0.537094. These are not fresh-confirmation intervals.

| Root backbone | Q0 Energy | Q1 Energy | Q2 Energy | Q2 gain [interval], % |
|---|---|---|---|---|
| CSDI | 0.555356 | 0.527114 | 0.525147 | 0.373 [0.126, 0.613] |
| TimeGrad | 1.195383 | 0.793076 | 0.783074 | 1.261 [0.813, 1.746] |
| D3U | 0.512819 | 0.511007 | 0.510180 | 0.162 [−0.081, 0.438] |
| TACTiS-2 | 0.523958 | 0.517718 | 0.514433 | 0.634 [0.521, 0.709] |

and worsens Energy by 10.774%. These are target-informed development results, not independent confirmation.

**Development attribution and mechanism comparisons.** The original development summary and detailed mechanism discussion below complement the locked-year results in Section 4.1.

Protection restricts the available actions, but does not make one executor unique. For graph contrasts $D$, choosing $U \in \mathrm{null}(D)$ protects those contrasts and leaves at most $K - \mathrm{rank}(D)$ active dimensions. At $K = 20$, ten disjoint pairs leave ten dimensions; a connected protected graph leaves only one. An independent SVD covariance projection reproduces this graph action. The appropriate test is therefore against matched feasible actions, not only against an uncorrected generator.

On GHCNh, the plugin improves an adapted CSDI by 3.38% Energy in both units, but scalar/static heads improve it more and marginal CRPS worsens (Appendix R). Solar channel-path permutations preserve marginal CRPS exactly, yet all three swap variants trail fixed simple references in Energy and both sum-CRPS measures across three backbones. Seven consecutive test days with three technical seeds are not 21 independent observations (Appendix T). Block-law Gram matching likewise has no resolved advantage over same-summary Energy (Appendix S).

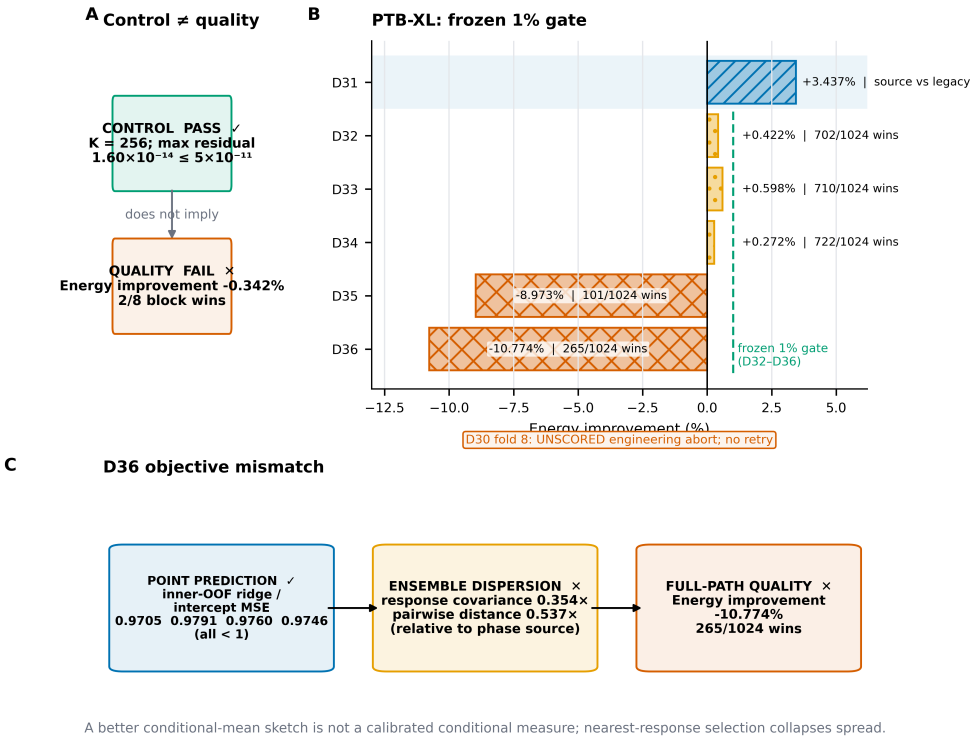


Figure 7: Control–quality separation and PTB-XL falsification. Exact OISST action coexists with a failed zero-request quality gate; the D36 point objective compresses ensemble dispersion and worsens full-path Energy.

Two falsifications show why the paper concerns generated laws rather than point accuracy. OISST rank-1 control passes below $1.6 \times 10^{-14}$, yet the unchanged generator worsens Energy by 0.342% and wins only 2/8 blocks (Huang et al., 2020). On consumed PTB-XL folds 1–7 (Wagner et al., 2020; 2022), response prediction improves pointwise error, but nearest-response path selection worsens Energy by 10.774% (265/1,024 wins). Covariance trace falls to 35.4% and pairwise distance to 53.7% of the anchor. The loss of joint-path quality accompanies collapsed spread. Neither point accuracy nor an exact action certificate can substitute for generation-quality evidence.

## H Proof of Theorem 1

Let $X, X' \sim Q$ and $Y, Y' \sim P$ be independent. The population Energy score risk of $Q$ under $P$ is

$$R(Q; P) = \mathbb{E}\|X - Y\|_2 - \tfrac{1}{2}\mathbb{E}\|X - X'\|_2.$$

Subtracting $R(P; P) = \frac{1}{2}\mathbb{E}\|Y - Y'\|_2$ gives

$$R(Q; P) - R(P; P) = \tfrac{1}{2}(2\mathbb{E}\|X - Y\|_2 - \mathbb{E}\|X - X'\|_2 - \mathbb{E}\|Y - Y'\|_2) = \tfrac{1}{2}D_{\mathrm{E}}(P, Q).$$

Energy distance is nonnegative and is zero only when the two distributions coincide under the usual finite first-moment condition. Hence a nonidentity request with $D_{\mathrm{E}}(P, Q) > 0$ has strictly larger expected Energy risk when it is scored against the unrequested law $P$.

## I Proof and boundary of Proposition 3

Since $R(C) \in O(r)$ and $T$ is invariant,

$$T(CR(C)) = T(C).$$

Moreover,

$$CQR(CQ) = CQQ^\top R(C)Q = CR(C)Q,$$

so the construction is gauge covariant rather than tied to a privileged thin basis. If the history measurement alone receives this right action while q0 and the target law remain fixed, a controller using the transformed covariant shape can take a different action. Proper risks can therefore differ although the gate inputs coincide.

Our finite construction reverses past-derived eigendirections and preserves the complete row Gram matrix. The reversed directional policy is worse than scalar in 32/32 replicates and changes every directional path; construction/controller gauge errors are $1.88 \times 10^{-14}$/$8.88 \times 10^{-15}$. The preregistered empirical conjunction does not pass because independently recomputed scalar paths are not literally bit-identical after floating-point transformation. Their maximum score-mean discrepancy is $5.56 \times 10^{-17}$. Thus the exact proposition is supported, while the formal empirical label remains unresolved.

## J Proof of Theorem 2

Because $U^\top U = I_r$, $(CG^\top - C)U^\top(I - UU^\top) = 0$, and left multiplication by $P_H$ preserves this equality. Since $\mathbf{1}^\top P_H = 0$, $\mathbf{1}^\top \Delta = 0$. Moreover $C = P_H XU$ is time centred, so

$$\begin{aligned} P_H X_G U &= P_H XU + P_H\{(CG^\top - C)U^\top\}U \\ &= C + (CG^\top - C) = CG^\top. \end{aligned}$$

The empirical covariance of the active rows therefore transforms as $\widehat{\Sigma}_{CG^\top} = G\widehat{\Sigma}_C G^\top$. For an orthogonal basis change $U' = UQ$ and conjugate request $G' = Q^\top GQ$,

$$(P_H XU'G'^\top - P_H XU')U'^\top = (CG^\top - C)U^\top,$$

so the physical action is gauge covariant. The identity branch bypasses arithmetic and returns the original array. Computing $XU$, the $r \times r$ request, and the reconstruction requires $O(NHKr + NHr^2)$ operations and thin intermediates.

## K Learning a protected predictive measure

Learning to correct a misspecified generator is compatible with the population separation result. The distinction is between a learned predictive law and an exogenous request. The Root-CSDI–TISS–paired-slack development stack makes this distinction operational. Its source is a support-native CSDI research variant, not a repaired version of the previously ineligible released-law comparator. TISS, a temporal-isometric set shift, learns a channelwise level correction. Paired-slack then learns redistribution within stated physical constraints. Together they produce a predictive measure before any external request is applied.

The algebra below states what each stage preserves. The completed development comparison then tests the quality of the resulting path distribution. Neither the algebra nor that comparison provides automatic qualification.

### K.1 The scoring object is part of the method

For any finite path ensemble $Z_{1:M}$, including members coupled by a learned set operation, define the issued probability measure $\widehat{Q}_Z = M^{-1}\sum_m \delta_{Z_m}$. Its Energy score is exactly

$$S_{\mathrm{V}}(Z, y) = \frac{1}{M}\sum_m \|Z_m - y\|_2 - \frac{1}{2M^2}\sum_{m,n}\|Z_m - Z_n\|_2 = S_{\mathrm{E}}(\widehat{Q}_Z, y). \tag{19}$$

This is not an estimator claimed to be unbiased for the marginal law of one member. In contrast, the U-statistic in Equation 1 estimates that population score without finite-ensemble bias under conditionally iid draws, or under another sampling scheme that provides the same required

pairwise expectations. Exchangeability alone is insufficient (Ferro, 2014). For $M \geq 2$, with $\overline{d}_{\neq} = [M(M-1)]^{-1} \sum_{m \neq n} \|Z_m - Z_n\|_2$, the exact difference is

$$S_{\mathrm{V}} - S_{\mathrm{U}} = \frac{\overline{d}_{\neq}}{2M}. \tag{20}$$

The difference depends on ensemble spread even at fixed $M$.

For a minimal counterexample, let a two-member ensemble be $(0, 2)$ or $(2, 0)$ with equal probabilities, and let the independent verification law place mass $1/2$ at both 0 and 2. The members are exchangeable and each has the correct marginal law. Nevertheless, the expected U-statistic is zero, while the population Energy risk is $1/2$. The issued empirical measure is exactly the verification law and its V-score risk is also $1/2$. This counterexample checks the sampling assumption; it is not a new scoring rule.

Scoring-rule training of generative forecasts is established prior work (Pacchiardi et al., 2024). Our distinction concerns the admissible transform and the object being scored. The newer coupled-ensemble development protocol uses Equation 19, including for its base arm, at common $M = 64$. Its declared, fit-frozen normalization is applied before evaluating the norm. The historical EEG and USGS off-diagonal-score decisions retain their frozen estimators. Their finite-ensemble comparative results do not by themselves establish the sampling assumptions required for an unbiased population-risk interpretation. These protocols and the newer finite-measure protocol remain separate; no historical score is recomputed or pooled across protocols.

### K.2 A COMPOSITION CERTIFICATE WITH EXPLICIT REFERENCE POINTS

Consider one context and a nonnegative physical ensemble $X \in \mathbb{R}_+^{M \times H \times K}$, with even $K$ and $M \geq 2$. TISS produces

$$Y_{m,t,k} = X_{m,t,k} + \delta_k, \qquad \delta_k \in [\ell_k, a], \quad \ell_k = \max\{-a, f_k - \min_{m,t} X_{m,t,k}\}. \tag{21}$$

Here $0 \leq f_k \leq \min_{m,t} X_{m,t,k}$ is a declared output floor and $a > 0$ is the physical action cap. The current development wrapper uses $f_k = 0.01 \min_{m,t} X_{m,t,k}$ and a cap of $0.25$ times the fit-frozen physical unit. Its lower shift bound is therefore $\max\{-a, -0.99 \min_{m,t} X_{m,t,k}\}$; the broader nonnegative ideal class has $f_k = 0$. The shift is constant over members and future times but may depend on the context and source ensemble.

For each time and declared channel pair $(2j-1, 2j)$, suppress $(t, j)$ and set $w_m = \min(Y_{m,2j-1}, Y_{m,2j})$, $R_m = Y_{m,\{2j-1,2j\}} - w_m(1,1)$ and $W = \sum_m w_m$. Given scores $|v_m| \leq b$, define

$$s_m = \begin{cases} W \dfrac{w_m e^{v_m}}{\sum_n w_n e^{v_n}}, & W > 0, \\ 0, & W = 0, \end{cases} \qquad Z_{m,\{2j-1,2j\}} = R_m + s_m(1,1). \tag{22}$$

The implementation uses $b = 2$. This is a standard exponential redistribution on a support face, not a new simplex optimization theorem.

**Proposition 1** (Reference-aware composition)**.** *In exact arithmetic, Equations 21–22 produce $Z \geq 0$ and preserve the following quantities:*

$$\begin{aligned} Y_{m,t,k} - \bar{Y}_{t,k} &= X_{m,t,k} - \bar{X}_{t,k}, \\ Y_{m,t+1,k} - Y_{m,t,k} &= X_{m,t+1,k} - X_{m,t,k}, \\ \bar{Z}_{t,k} &= \bar{Y}_{t,k} = \bar{X}_{t,k} + \delta_k, \\ Z_{m,t,2j-1} - Z_{m,t,2j} &= X_{m,t,2j-1} - X_{m,t,2j} + \delta_{2j-1} - \delta_{2j}. \end{aligned} \tag{23}$$

*Bars denote uniform member means. On each positive-slack coordinate, $e^{-2b} \leq s_m/w_m \leq e^{2b}$; zero-slack coordinates remain zero. Consequently, $Z \geq e^{-2b} Y$ coordinatewise.*

*Proof.* The lower shift bound makes $Y$ nonnegative. Subtracting the same $\delta_k$ across members or times proves the first two identities. The residual $R_m$ is nonnegative with at least one zero entry. Equation 22 gives $s_m \geq 0$ and $\sum_m s_m = W$, so member sums remain unchanged in both channels. Adding the same $s_m - w_m$ to the two channels preserves their difference relative to $Y$, giving the

last two identities. For $W > 0$, the denominator lies in $[e^{-b}W, e^{b}W]$, which yields the ratio bounds. Since $R_m \geq 0$ and $e^{-2b} \leq 1$, $R_m + s_m(1,1) \geq e^{-2b}\{R_m + w_m(1,1)\}$. The $W = 0$ case is immediate. □

Protection must be read relative to the stated input. The last two identities protect physical means and pair differences relative to $Y$, not all properties relative to $X$. TISS can change a pair difference through unequal channel shifts. Paired-slack can change member anomalies and temporal increments because $s_m - w_m$ varies with member and time. Neither the full protected complement, cumulative-endpoint invariance, global invertibility, nor a non-collapse guarantee follows for this composition. Physical mean preservation also does not imply mean preservation after a nonlinear log transform. Any subsequent external KRPC request requires its own compatible frame and support checks. It is not an automatically certified fourth stage.

For exactly the stated pair and mean constraints, the feasible $s$ values form the scaled simplex $\{s \geq 0 : \sum_m s_m = W\}$. Positive, unbounded logits cover its interior only when every $w_m > 0$. In the implemented bounded-logit, shared-network class, neither all simplex points nor arbitrary independent logits are available. Member-equivariant features make the transport insensitive to member labeling, but do not make output members independent. Identical members receive identical scores and cannot acquire distinct paths from this deterministic map alone.

### K.3 A DIAGNOSTIC CEILING FOR CENTRE CORRECTION

The composition also bounds what further control capacity can change. Let $\mu^*_{t,k}$ be any declared desired mean profile in physical space and let $e_{t,k} = \mu^*_{t,k} - \bar{X}_{t,k}$ and $\bar{e}_k = H^{-1}\sum_t e_{t,k}$. This profile may be an explicit request or a fit-internal diagnostic; it is not assumed to be the unknown true conditional mean or a future observation.

**Proposition 2** (Ideal attainable-mean projection)**.** *Over the closed ideal shift class in Equation 21, followed by any redistribution preserving $\bar{Y}$, the smallest squared discrepancy from $\mu^*$ is*

$$\min_{\delta_k \in [\ell_k, a]} \sum_{t,k} (\mu^*_{t,k} - \bar{Z}_{t,k})^2 = \sum_{t,k} (e_{t,k} - \bar{e}_k)^2 + H \sum_k \operatorname{dist}(\bar{e}_k, [\ell_k, a])^2. \tag{24}$$

*It is attained in this ideal class by the interval projection $\delta^*_k = \operatorname{proj}_{[\ell_k, a]}(\bar{e}_k)$.*

*Proof.* Equation 23 reduces the discrepancy to $\sum_{t,k}(e_{t,k} - \delta_k)^2$. For each channel, $\sum_t (e_{t,k} - \bar{e}_k) = 0$, so orthogonal expansion gives $\sum_t (e_{t,k} - \bar{e}_k)^2 + H(\bar{e}_k - \delta_k)^2$. Minimizing the second term on its interval proves the result. □

The first term cannot be removed by a constant-in-time shift, even with a larger cap. The second measures the amplitude or output-floor restriction. A finite shared neural network may realize only a subset of this ideal box; a tanh endpoint may be approached only as a limit. The expression is therefore a lower bound for that network, not a promise that training attains it. It is an elementary projection identity used to explain the architecture's limitation, not a novel optimization theorem. It is also not an Energy-risk bound, a point-prediction leaderboard, or proof of poor conditional generation. Increasing paired-slack capacity cannot change either term while its mean constraint remains fixed.

### K.4 LEARNING UNDER MISSPECIFICATION

For a true conditional law $P$, a base forecast $Q_0$ and a learned forecast $Q_\theta$, all with finite first moments, the existing Energy identity gives

$$R(Q_\theta; P) - R(Q_0; P) = \tfrac{1}{2}\{D_{\mathrm{E}}(P, Q_\theta) - D_{\mathrm{E}}(P, Q_0)\}. \tag{25}$$

Thus a nonidentity learned correction can help when $Q_0 \neq P$; support and protection alone do not ensure it helps. For example, if $P = (\delta_0 + \delta_2)/2$ and $Q_0 = (\delta_1 + \delta_3)/2$, shifting by $-1$ reduces risk from 1 to $1/2$. This is calibration, not evidence that an arbitrary external request improves an already correct predictive law. For coupled ensembles, the same identity can be applied to the issued finite measures and then averaged over their construction randomness, provided that randomness

is independent of the verifying future conditional on the observed history and the expected first moments are finite. This does not require the members to be mutually independent.

These propositions supply a method-level bridge and falsifiable capacity limits. They do not establish superiority of exponential redistribution over a matched ray. Qualification on untouched outcomes and fair modern-baseline comparison remain separate empirical requirements.

### K.5 The peak that a protected residual cannot remove

For one time and channel pair, write $d_m = Y_{m,1} - Y_{m,2}$, $r_m = |d_m|$, and $W = \sum_m \min(Y_{m,1}, Y_{m,2})$. Preserving each member's difference and both channel means, while maintaining nonnegativity, requires $Z_m = ((d_m)_+, (-d_m)_+) + s_m(1,1)$ with $s_m \geq 0$ and $\sum_m s_m = W$. In this relaxed full-simplex class, the smallest possible pair peak is

$$\min_{s \geq 0,\, \sum_m s_m = W} \max_{m,k \in \{1,2\}} Z_{m,k} = \tau \equiv \max\left\{ \max_m r_m, \frac{\sum_m r_m + W}{M} \right\}. \tag{26}$$

The residual forces the first bound. Summing $r_m + s_m \leq \tau$ forces the second. Conversely, capacities $\tau - r_m$ are nonnegative and sum to at least $W$, so a slack allocation within these capacities attains the bound. Independent allocation across times and pairs gives the global peak floor by taking their maximum. This elementary minimax identity describes the constraint class, not a distribution-quality optimum or a new optimizer.

For bounded-entropy logits $|v_m| \leq b$, an individual slack coordinate has the tighter minimum

$$s_m \geq \frac{W w_m}{w_m + e^{2b}(W - w_m)}, \qquad W > 0. \tag{27}$$

It follows by taking $v_m = -b$ and every other logit to $b$; $W = 0$ gives zero. These individual extrema need not be attained simultaneously. Their largest residual-plus-slack value, combined with Equation 26, is therefore only a necessary peak lower bound for the implemented network.

### K.6 Transfer of control gains to a joint diffusion base

To test whether constrained corrections help a joint diffusion generator, we evaluated the fixed support-native Root-CSDI stack on 192 MERRA-2 development origins from 2020, with $K = 128$, $H = 16$ and $M = 64$. One 2018-only backbone and one 2019 training bank were shared by three adapter initializations. Root-CSDI and the NativeRootAR anchor each supplied one ensemble per origin. The three adapter stages supplied nine additional outputs, giving 2,112 nonduplicated records but only 192 time-origin units. Models were frozen before this evaluation. The year had already informed earlier development; the analysis is not fresh confirmation. Table 22 reports all eight prespecified contrasts. Adapter scores are first averaged within origin; circular block length 16 and 16,384 shared resamples yield Bonferroni-adjusted 99.375% conditional intervals. They do not remove historical selection or single-backbone uncertainty.

Table 22: Complete same-backbone development contrasts. Improvement is $100(1 - \text{candidate}/\text{base})$; positive values favor the candidate. All intervals are conditional 99.375% intervals, adjusted across eight comparisons. Q1 denotes TISS; Q2-O and Q2-S denote original- and structure-objective paired-slack. VS is the exact spatial variogram score.

| Candidate / base | Metric | Improvement (%) | Interval (%) |
|---|---|---|---|
| Q1 / Root-CSDI | Energy V | 2.13160 | [0.96901, 3.41865] |
| Q2-O / Q1 | Energy V | 0.81685 | [0.52366, 1.07306] |
| Q2-S / Q1 | Energy V | 0.80428 | [0.53438, 1.03604] |
| Q2-S / Q2-O | Spatial VS | 0.02888 | [−0.01168, 0.07198] |
| Root-CSDI / RootAR | Energy V | −3.40018 | [−15.28198, 4.04047] |
| Q2-S / RootAR | Energy V | −0.38219 | [−10.72709, 5.90170] |
| Q2-S / Q1 | CRPS V | 0.05287 | [−0.25538, 0.32876] |
| Q2-S / Q1 | Spatial VS | 0.56641 | [0.23441, 0.88472] |

Both control objectives improved Energy beyond TISS on this joint diffusion base. The reductions were positive for all three adapter initializations and four contiguous time bins. Q2-S also improved

spatial VS over Q1, but its extra spatial advantage over Q2-O was not established. Neither the base nor the full Q2-S stack established Energy superiority over RootAR; their mean scores were respectively 0.555356, 0.539147 and 0.537094 for RootAR. Intervals crossing zero do not establish equality or noninferiority. The full Q2-S stack reduced mean Energy by 2.91874% relative to its own base, a descriptive contrast outside the eight interval-tested comparisons.

The gains incurred identifiable costs. Relative to Q1, Q2-S increased mean tail coexceedance Brier by 1.38160% and the short-window normalized q99 absolute discrepancy by 25.09032%. Its spread increased 16.94082% and effective rank decreased 0.77346%; increment Energy improved 0.60653%. These are descriptive changes, not additional confirmatory tests. All outputs retained 64 distinct members and the stated physical constraints, with maximum mean and pair errors of $8.89 \times 10^{-15}$ and $2.85 \times 10^{-14}$. Those checks do not certify non-copying or complete temporal dependence preservation. Only 2.86458% of Q1 channel actions were near their amplitude limit on average, so a single maximum shift cannot establish pervasive saturation.

### K.7 Saved-sample capacity diagnosis

The saved ensembles distinguish a constraint on removing existing peaks from newly learned amplification. After the complete comparison, we evaluated Equations 26–27 on all 576 saved Q1 origin–initialization ensembles. This post-hoc analysis read generated outputs only: it neither accessed verification targets nor rescored quality. The ideal simplex peak floor retained a median 95.48638% of the input peak; the bounded-entropy necessary floor retained 95.57061%. Thus the corresponding median ceilings on peak reduction were 4.51362% and 4.42939%, not promised improvements. The observed Q2-S peak instead increased by a median 14.74939%, while its zero-score identity action was feasible. The protected residual limits removal of input peaks, but does not force this additional amplification.

This diagnosis separates two interventions. Reducing peaks already encoded in protected pair differences requires an upstream change or a different declared protection contract. Avoiding newly learned amplification is an objective and optimization question within the current class. A large peak is not intrinsically incorrect, and peak reduction is not a proper-score objective. These bounds do not prove that either change will improve generation. The following development experiment tests these two interventions while retaining the present models and results. No 2021–2025 qualification or outer targets were opened.

### K.8 Calibration capacity and tail-objective ablation

We tested whether upstream scale calibration or tail-aware learning could improve the remaining quality tradeoff. This experiment reused the fixed Root-CSDI backbone and 2019 ensemble bank. A date-only partition supplied 191 January–September fit origins and 59 year-end guard origins; six intermediate origins were excluded by the temporal separation rule. The earliest guard context began 120 hours after the fit cutoff. All 18 adapter terminals were frozen before guard parsing. The experiment completed 6,600 updates and 1,121 evaluation records in 79.43 CPU minutes. One backbone and one bank were shared by three adapter initializations; historical exposure to 2019 and 2020 remains an adaptive-development limitation.

The calibration factor compared shift-only Q1 (S) with positive affine Q1 (A). In physical-unit coordinates, the latter applies $Y_{mtk} = g_k X_{mtk} + b_k$, with $g_k = \exp(\log(2)\tanh(a_k))$ and $b_k = \max\{.25\tanh(c_k), (.01 - g_k)\min_{m,t} X_{mtk}\}$. Gain and shift are shared across members and forecast times. The affine model adds 33 active parameters to the 4,801-parameter shift model. It preserves within-channel order and standardized shape, not exact amplitudes. Each Q2 protects its own calibrated input's physical mean and per-member paired-channel difference. These are not end-to-end Q0 invariants.

For each calibration, we compared the structure objective (S2) with a tail-augmented objective (T2). Both use the same 161-parameter Q2, initialization, minibatch tape and 400-update budget; each Q1 receives 300 updates. Let $z$ denote the frozen standardized log coordinate and $u$ the transformed 2018 channelwise 95th-percentile threshold. We apply $v(z) = .25\,\mathrm{softplus}((z - u)/.25)$ to both forecasts and observations. The tail score is Energy-V plus .25 CRPS-V in these transformed coordinates, following the existing transformed-kernel-score framework (Allen et al., 2023). The added

term is weighted by $.25c_T$; $c_T$ is the ratio of the mean structure objective to mean tail score on the 191 unmodified Q0 fit ensembles. The spatial coefficient is likewise fixed from Q0. All modes and initializations share these coefficients. No weight, budget or checkpoint is selected on the guard.

Table 23: All eight scale–tail simple-effect comparisons. S/A denote shift/affine calibration; S2/T2 denote structure/tail-augmented Q2. Tail denotes the transformed score, not raw-tail Brier. Positive relative reductions favor the candidate. Intervals are conditional 99.375% intervals.

| Candidate / base | Metric | Reduction (%) | Interval (%) |
|---|---|---|---|
| A-Q1 / S-Q1 | Energy V | 0.06983 | $[-1.25223, 1.40448]$ |
| A-Q1 / S-Q1 | Tail | 1.48618 | $[-1.75029, 4.09094]$ |
| A-S2 / S-S2 | Energy V | $-0.27647$ | $[-1.55390, 1.00180]$ |
| A-S2 / S-S2 | Tail | 0.51957 | $[-2.51303, 3.00487]$ |
| S-T2 / S-S2 | Energy V | 0.03305 | $[-0.04760, 0.09742]$ |
| S-T2 / S-S2 | Tail | 0.20455 | $[-0.04302, 0.36621]$ |
| A-T2 / A-S2 | Energy V | 0.03016 | $[-0.01377, 0.06809]$ |
| A-T2 / A-S2 | Tail | 0.08889 | $[-0.02832, 0.16578]$ |

Scores are averaged across adapter initializations within each origin. Circular blocks of eight origins and 16,384 shared bootstrap resamples produce the Bonferroni-eight intervals in Table 23. All eight intervals cross zero. The comparisons do not establish superiority, equivalence or noninferiority. They test prespecified simple effects, not a factorial interaction. Only 59 late-year time units are available, approximately 7.4 nominal blocks; the intervals do not account for historical selection or backbone uncertainty.

Descriptive tradeoffs also limit the case for the added components. Affine Q1 increased raw-tail Brier by 3.18213% and marginal CRPS by 1.05086% relative to shift Q1. Under shift calibration, the added tail term reduced its transformed score but increased marginal CRPS by 0.07736%, spatial VS by 0.07065%, raw-tail Brier by 0.06026%, and short-window q99 discrepancy by 3.10497%. Spread increased 1.89170% and effective rank decreased 0.49835%. These are descriptive changes, not additional interval-tested comparisons. All outputs retained 64 distinct members and the declared support; the largest Q2 mean and pair errors were both $1.78 \times 10^{-15}$.

Neither intervention established a robust improvement over its matched alternative. The earlier same-backbone control gain remains, and the available headroom is not proved exhausted. Admissible control, larger calibration capacity and a lower tail-weighted mean score therefore still require a complete distribution-quality comparison. We retain these results without selecting a new winner or opening the 2021–2025 evaluation targets.

**Training response at the fixed terminal models.** After seeing the complete guard results, we evaluated all 18 fixed models on all 191 original fit origins, without updates or guard rescoring. Affine Q1 reduced fit Energy by 3.60159% and transformed-tail score by 7.58763% relative to shift Q1. Both directions held for all three initializations. The corresponding guard changes were only 0.06983% and 1.48618%, with intervals crossing zero. Affine calibration therefore changed the learned distribution and improved these empirical fit scores. These fit gains do not establish temporal generalization. The result is not evidence that the calibration pathway was inactive.

The tail-objective intervention had a different response. Its transformed fit score improved by 0.22030% under shift calibration and 0.08617% under affine calibration, while fit Energy worsened by 0.01243% and 0.01290%, respectively. Even the combined tail objective improved by only 0.01776% and 0.00107% on average; one of three shift terminals and two of three affine terminals were worse than their structure-trained counterparts on that same combined objective. These descriptive checks do not establish that empirical optimization is complete. Affine gain medians ranged from 1.035 to 1.046 across initializations, with 95th percentiles of absolute log gain from 0.096 to 0.108. A few gains approached the lower bound, but a single minimum does not establish widespread saturation. This post-hoc diagnosis separates fit response from generalization; it does not identify overfitting or temporal shift as a unique cause, and does not select a replacement checkpoint.

### K.9 TEMPORAL TRANSFER OF CALIBRATION AND STABILITY REGULARIZATION

**Question and temporal design.** The preceding year-end comparison left temporal transfer unresolved despite a measurable fit response. Experiment C20260909b examined this transfer with an expanding-window design rather than another year-end weight search. It reused the 2018-only Root-CSDI checkpoint and the existing 2019 $M = 64$ bank at $K = 128$, context length 96, and horizon 16. Fold 0 fitted bank indices 0–63 and evaluated 70–126; fold 1 fitted 0–127 and evaluated 134–190. Each evaluation contained 57 origins, with full raw-window embargoes of 124 and 125 hours, respectively. The later model can train on dates used to evaluate the earlier, already frozen model. These are therefore dependent expanding-window folds, not independent replications. The historically exposed 2019 data remain adaptive development data. The old year-end guard and 2020–2025 were not reopened.

**Matched interventions.** We compared shift ERM (S), positive-affine ERM (A), and positive-affine stability regularization (AS). Each used 300 Q1 calibration updates followed by 400 Q2 updates with the same paired-slack architecture and spatially augmented objective. The two affine arms share capacity; shift and affine have 4,801 and 4,834 active Q1 parameters, respectively. Initialization and the stratified minibatch tape are matched within each fold and replicate. Two folds, three replicates, and three arms give 12,600 updates and 36 terminal models. All terminals were completed before evaluation. The single CPU run finished in 4.78 hours without restart or checkpoint selection. Its 114 origins produced 2,166 complete records: one Q0 and 18 Q1/Q2 outputs per origin. There was no backbone retraining or new proposal sampling.

**What the stability term penalizes.** Using the fixed 2018 log-standardization, write the Q1 loss as

$$\ell_0 = \mathrm{Energy}_V + 0.25\,\mathrm{IncrementEnergy}_V + 0.25\,\mathrm{CRPS}_V.$$

Let $\mu$ be the full-fold Q0 mean loss. Four equal chronological fit blocks each contribute two independent draws with replacement. With $\widehat{d}_b$ the block minibatch mean of $D_{bi} = (\ell_{0,\theta,bi} - \ell_{0,Q0,bi})/\mu$, AS minimizes the ordinary minibatch loss plus $0.25\mu R$, where $R = \frac{1}{4}\sum_{b=1}^{4}(\widehat{d}_b - \overline{\widehat{d}})^2$. Risk-variance regularization has prior foundations in REx (Krueger et al., 2021). This paired, finite-bank objective is not presented as a new general regularization principle or an exact replication of V-REx.

For fixed parameters and the fixed finite fit bank, let $m_b$ and $\sigma_b^2$ be the block mean and population-denominator variance of $D_{bi}$. Independent index sampling gives the exact identity

$$\mathbb{E}[R] = \mathrm{Var}_b(m_b) + \frac{3}{8}\,\mathrm{mean}_b(\sigma_b^2). \tag{28}$$

Indeed, expand $R$ as the mean squared block estimate minus the squared grand mean. Each block estimate has variance $\sigma_b^2/2$, and cross-block sampling covariance is zero. Subtraction yields the coefficient $3/8$. This argument concerns random indices conditional on a finite bank; it does not assume naturally occurring time windows are independent. Exhaustive enumeration of 256 synthetic draw combinations verified both values and raw gradients at differentiable points. It does not establish unbiased clipped-Adam parameter updates. Both terms measure variation in paired loss changes, not physical path dispersion. Identity, uniform improvement, and uniform deterioration all have zero penalty. A small penalty is thus neither a quality guarantee nor evidence of diverse paths.

| Fold | Candidate / base | Energy-V | CRPS-V | Spatial VS | Raw-tail Brier |
|---|---|---|---|---|---|
| 0 | A-Q1 / S-Q1 | 0.34995 | -0.22494 | 0.41751 | -1.52788 |
| 1 | A-Q1 / S-Q1 | 7.20444 | 1.81693 | 4.79103 | 1.17104 |
| 0 | A-Q2 / S-Q2 | 0.06938 | -0.25797 | 0.21727 | -1.58068 |
| 1 | A-Q2 / S-Q2 | 6.85099 | 1.77048 | 4.47360 | 1.19781 |
| 0 | AS-Q1 / A-Q1 | 0.00098 | 0.00677 | 0.00177 | 0.01699 |
| 1 | AS-Q1 / A-Q1 | -0.00853 | -0.00112 | -0.00543 | -0.00772 |
| 0 | AS-Q2 / A-Q2 | 0.00115 | 0.00544 | 0.00256 | 0.01725 |
| 1 | AS-Q2 / A-Q2 | -0.00860 | -0.00154 | -0.00504 | -0.00558 |

Table 24: All eight prespecified C20260909b fold-level comparisons. Entries are percentage reductions in replicate-averaged mean scores; positive values favor the candidate. These are descriptive contrasts, without confidence intervals, hypothesis tests, or an adjustment that would remove historical adaptivity. All 14 metrics and all replicates are retained in the complete result ledger.

**Calibration transfers unevenly.** Table 24 shows larger affine gains in fold 1 than fold 0. The three Q1 Energy reductions span 6.95401–7.49078% in fold 1 and 0.25367–0.50948% in fold 0. Full-stack Energy reductions are 6.85099% and 0.06938%, respectively. The earlier fold incurs CRPS and raw-tail costs. Fit size, fit dates, and evaluation dates all change between folds, so this heterogeneity cannot identify seasonality as its unique cause. It also does not negate the preceding year-end null result.

**The stability penalty has not yielded a consistent quality gain.** AS versus A changes full-stack Energy by $+0.00115\%$ and $-0.00860\%$. Saved full-fit block means show a 0.31–0.70% reduction in between-block dispersion across the six matched Q1 terminals. That response measures only the between-block component, not the entire expectation in Equation 28. It has not translated into a consequential, consistent evaluation gain. These results establish neither equivalence nor adequate optimization. They do not show exhausted headroom or failure of all stability methods. No weight, budget, or terminal was changed after observing these results.

**Separate calibration and control contributions.** Additional post-hoc arithmetic from the complete saved means shows that A-Q2 reduces Energy relative to its own A-Q1 input by 0.45285% and 0.57328%. Its raw-tail Brier improves by 0.66981% in fold 0 but worsens by 0.78381% in fold 1. The full A-Q2 stack improves Energy over Q0 by 1.50581% and 9.29120%; these full-stack gains must not be attributed exclusively to Q2 or KRPC. These attribution contrasts have no new significance tests. Diversity is also not uniformly improved: A-Q2 versus S-Q2 reduces effective rank by 0.75408% and 0.67394%, while mean pairwise spread increases by 0.88416% in fold 0 and decreases by 0.30596% in fold 1. Neither direction of spread alone establishes quality or collapse.

**Verification and evidence boundary.** All 36 terminal models passed strict CPU reload with finite parameters and optimizer state. Reaggregation of all 2,166 unique records exactly reproduced the saved summary, with 30,324 finite metric values. Across 1,026 Q2 certificates, the maximum mean and paired-difference errors were $7.11 \times 10^{-15}$ and $1.14 \times 10^{-13}$; the minimum support margin was $2.37 \times 10^{-6}$. Q2 protects its own Q1 input, not every Q0 invariant. All ensembles retained 64 distinct members; zero matches against the finite reference bank is not a universal non-memorization guarantee. The next scientific bottleneck is a support-matched comparison with strong modern generators and a separately frozen independent evaluation, not another retrospective regularization sweep. This experiment provides adaptive temporal generation evidence, not an all-roster SOTA result. MERRA-2 2021–2025 were sealed at this experiment's closeout; the subsequent studies have their own access history.

## L MATCHED MODERN-BACKBONE GENERATION QUALITY

**Separating proposal, calibration and protected correction.** The comparison follows three issued ensembles: a support-native proposal Q0, generic positive-affine calibration Q1, and entropy paired-slack correction Q2 of that Q1. This separates the quality of the underlying generator from the gains supplied by calibration and constrained correction. Here Q1 means affine calibration, not the earlier shift-only TISS in Appendix K.6. The comparison includes Root-CSDI, Root-TimeGrad, Root-D3U and Root-TACTiS-2, based on the published model families (Tashiro et al., 2021; Rasul et al., 2021; Li et al., 2025; Ashok et al., 2024), plus NativeRootAR. The square-root training coordinates and CPU float64 square pushforward were specified for these distinct variants before their fits. They are not repairs to previously ineligible log-law outputs, nor claims to reproduce the strongest published configuration. Root-D3U uses the PatchTST/PatchDN conditional-product variant, not a learned joint copula or the SparseVQ configuration.

**Matched adaptation and unequal backbone budgets.** All neural Q0 fits use 2018 observations. Each architecture supplies one M64 ensemble at each of the same 256 past-only 2019 origins for adaptation. Three adapter initializations share that architecture's bank. Q1 has 4,834 parameters and 300 AdamW updates; Q2 has 161 parameters and 400 Adam updates. Capacity, initialization and minibatch tapes, losses and adaptation calendars are matched across the four backbones. The base training budgets are not: CSDI uses 5,000 updates with batch 8, TimeGrad 5,000 with batch 64, and D3U 2,710 point-model plus 27,100 residual-model updates with batch 32. TACTiS-2 uses 2,500 marginal-flow and 2,500 copula updates, with compact capacity and 20-channel training bags

followed by full-$K$ joint sampling. Accordingly, this design matches adaptation opportunity, not training FLOPs or optimized architecture performance. Three adapters conditional on one base fit are not three independent backbone replications.

Q1 learns a bounded positive gain and a support-safe shift. Q2 preserves its own Q1 ensemble mean over members, the declared adjacent-channel pair differences, and a 1% support floor. The shared base loss combines normalized log1p Energy V, increment Energy V and marginal CRPS V; Q2 additionally uses the frozen spatial-variogram term. These constraints do not preserve every temporal or distributional property. The paired-slack class is distinct from the thin-SPD request compiler and does not inherit its endpoint, protected complement or gauge certificates automatically.

**Evaluation and statistical unit.** All scores use the same 192 origins from the previously exposed year 2020, with context 96, horizon 16, $K = 128$ and $M = 64$. Quality coordinates are log1p of physical values, normalized using 2018, not the square-root training coordinates or log1p of unit-divided values. Energy V scores the issued finite measure with denominator $2M^2$, as justified in Appendix K. Ensemble coupling does not authorize an iid fair-U correction. Thirteen declared diagnostics cover joint and marginal law quality, dynamic increments, spatial dependence, raw tails, diversity and finite-reference copying. Pointwise MAE does not select the preferred generator.

The D3U extension contributes 1,344 new records and reuses 6,336 unique legacy records without redrawing or rescoring them. The resulting 7,680 records cover 16 configurations, including historical TISS and original-slack arms. The subsequent TACTiS-2 extension adds 1,344 records, reusing those 7,680 without new legacy draws or scores: 9,024 unique records across 19 configurations. Table 25 shows all thirteen current affine Q0/Q1/Q2 or anchor configurations; the historical configurations remain in their original result ledgers and reports. Statistical replication is at the 192-origin level, not the record, member, channel or adapter level. Adapter scores are averaged within origin before circular block-16 resampling with 16,384 draws. Table 26 retains all twelve D3U contrasts and their original Bonferroni-adjusted 99.583333% conditional intervals. Table 27 retains all fifteen TACTiS-2 contrasts at their original 99.666667% conditional level. This revision neither pools historical families nor changes their intervals. The adjustment does not remove adaptive reuse of 2020, single-backbone uncertainty or dependence extending beyond the block length.

Table 25: Completed support-matched generation development panel. All five listed scores are lower-is-better. Q1 is affine calibration; Q2 is its structure-objective paired-slack correction. Values average adapter initializations within each origin, then origins. Descriptive rankings do not establish superiority. The remaining eight diagnostics and historical configurations are retained in the saved report.

| Configuration | Energy V | CRPS V | Increment | Spatial VS | Tail Brier |
|---|---|---|---|---|---|
| Root-CSDI Q0 | 0.555356 | 0.318275 | 0.159145 | 0.135043 | 0.007744 |
| Root-CSDI Q1 | 0.527114 | 0.313968 | 0.158237 | 0.130247 | 0.007376 |
| Root-CSDI Q2 | 0.525147 | 0.313449 | 0.157653 | 0.129820 | 0.007428 |
| Root-TimeGrad Q0 | 1.195383 | 0.656143 | 0.289994 | 0.332043 | 0.012239 |
| Root-TimeGrad Q1 | 0.793076 | 0.443163 | 0.270412 | 0.210546 | 0.009799 |
| Root-TimeGrad Q2 | 0.783074 | 0.440889 | 0.276267 | 0.208510 | 0.009610 |
| Root-D3U Q0 | 0.512819 | 0.314396 | 0.157585 | 0.130906 | 0.007125 |
| Root-D3U Q1 | 0.511007 | 0.313087 | 0.157795 | 0.130867 | 0.007039 |
| Root-D3U Q2 | 0.510180 | 0.312054 | 0.157294 | 0.130112 | 0.006989 |
| Root-TACTiS-2 Q0 | 0.523958 | 0.316299 | 0.274788 | 0.133803 | 0.006868 |
| Root-TACTiS-2 Q1 | 0.517718 | 0.315108 | 0.267495 | 0.132068 | 0.006816 |
| Root-TACTiS-2 Q2 | 0.514433 | 0.313611 | 0.278496 | 0.131339 | 0.006794 |
| NativeRootAR | 0.537094 | 0.334301 | 0.156618 | 0.140594 | 0.007396 |

**Calibration headroom and the additional control gain.** In the preceding fourteen-contrast matched-affine study, TimeGrad's affine Q1 improved Energy over its TISS by 27.45469% [23.67646%, 30.67689%]. Its affine-Q2 over affine-Q1 Energy improvement was 1.26125% [0.81251%, 1.74638%]; CSDI's was 0.37310% [0.12579%, 0.61293%]. Those intervals retain their original 99.642857% level and are not part of the twelve-comparison D3U family. TimeGrad's increment Energy instead worsened by 2.16528%, with a conditional deterioration interval of [1.03306%, 3.45557%] and an adverse direction in all four time bins. CSDI's short-window q99 discrepancy

Table 26: All twelve predeclared D3U extension contrasts. Improvement is $100(1 - \text{candidate}/\text{reference})$; positive values favor the candidate. Intervals are the original 99.583333% conditional development intervals. An interval crossing zero establishes neither equality nor noninferiority. Q1 and Q2 have the meanings in Table 25.

| Candidate / reference | Metric | Improvement (%) | Interval (%) |
|---|---|---|---|
| D3U Q1 / D3U Q0 | Energy V | 0.35334 | [−0.01320, 0.78709] |
| D3U Q2 / D3U Q1 | Energy V | 0.16178 | [−0.08141, 0.43815] |
| D3U Q2 / D3U Q1 | Increment Energy V | 0.31775 | [0.08428, 0.61187] |
| D3U Q2 / D3U Q1 | Spatial VS | 0.57701 | [0.24066, 0.98353] |
| D3U Q2 / D3U Q1 | Tail Brier | 0.69916 | [−0.60718, 2.25925] |
| D3U Q2 / RootAR | Energy V | 5.01098 | [3.40455, 7.00073] |
| D3U Q0 / CSDI Q0 | Energy V | 7.65944 | [1.56237, 16.69169] |
| D3U Q0 / TimeGrad Q0 | Energy V | 57.10002 | [53.22199, 60.87575] |
| D3U Q1 / CSDI Q1 | Energy V | 3.05564 | [−0.20039, 8.54384] |
| D3U Q2 / CSDI Q2 | Energy V | 2.85001 | [−0.28994, 8.40902] |
| D3U Q1 / TimeGrad Q1 | Energy V | 35.56647 | [30.98567, 40.49404] |
| D3U Q2 / TimeGrad Q2 | Energy V | 34.84899 | [30.30993, 39.77738] |

worsened descriptively by 16.99886%. A control certificate cannot cancel these quality costs. The large TimeGrad calibration gain establishes unused capacity in the prior TISS comparison, not saturation of the affine class.

D3U's full stack has the lowest mean Energy in the completed panel, but its advantage over CSDI Q2, the nearest competitor in the original three-backbone study, is not established by the retained interval. TACTiS-2 Q2 is now the nearest other-backbone Q2 by mean Energy; its declared comparison with D3U Q2 also has an interval crossing zero. D3U's own Q2-over-Q1 Energy interval crosses zero. Increment and spatial effects have positive conditional intervals; these support property-specific development gains. The 5.01098% full-stack advantage over RootAR is not a 5.01098% control contribution, since D3U Q0 already supplies the stronger Energy starting point. RootAR still has lower mean increment Energy, and CSDI Q2 lower mean spatial VS, than D3U Q2. Even the lowest-Energy stack therefore does not dominate every listed property.

**Retained adverse and non-ranking diagnostics.** D3U Q2 worsens Energy by 0.12906% and increment Energy by 0.00844% in the third contiguous time bin, despite the pooled positive means. Its tail Brier improves in the first bin but worsens in the remaining three; the pooled 0.69916% improvement has an interval crossing zero. Lag-1 and lag-16 correlations of that difference are 0.7070 and 0.2142, respectively. The q99 discrepancy improves by 6.85271% over its own Q1 but remains 9.11993% worse than CSDI Q2. These q99 differences are descriptive, not extra interval-tested comparisons or proper scores. Spread and effective rank decrease by 8.70867% and 2.33825%; neither quantity has a universal preferred direction. All 576 D3U Q2 origin–adapter certificates pass, with maximum mean and pair errors $5.33 \times 10^{-15}$ and $3.55 \times 10^{-15}$. All issued ensembles retain 64 distinct members and zero matches against the finite reference bank. Those checks establish neither complete non-copying nor privacy, nor preservation of the entire joint law.

**The copula extension sharpens the tradeoff.** TACTiS-2 Q1 improves whole-path Energy over Q0 by 1.19092%; Q2 adds 0.63446% over its own Q1. The latter has positive effects in all three adapter initializations and all four time bins. Its spatial VS improves by 0.55218%. Yet increment Energy deteriorates by 4.11267%, with an adverse direction in every initialization and time bin. The increment-difference lag-1 and lag-16 correlations are 0.7348 and 0.3736: dependence beyond the chosen block remains a limitation, not grounds for a post hoc block change. Tail Brier's 0.32078% improvement has an interval crossing zero, while short-window q99 discrepancy increases by 8.32612% descriptively. Thus an improved threshold-exceedance score does not establish amplitude-tail fidelity.

The TACTiS-2 correction preserves its own Q1 member mean and adjacent-channel pair differences within $8.88 \times 10^{-16}$ and $1.78 \times 10^{-15}$, respectively; all 576 origin–adapter certificates pass. All ensembles remain finite, nonnegative and 64-distinct, with no matches to the finite reference bank. These certificates coexist with the temporal cost. The square map folds 61 negative latent-root values

Table 27: All fifteen predeclared TACTiS-2 extension contrasts. Positive improvement favors the candidate; all intervals are the original Bonferroni-15 99.666667% conditional development intervals. Negative increment effects are retained, not offset by Energy gains. T2 denotes Root-TACTiS-2.

| Candidate / reference | Metric | Improvement (%) | Interval (%) |
|---|---|---|---|
| T2 Q1 / T2 Q0 | Energy V | 1.19092 | [0.15956, 2.37751] |
| T2 Q2 / T2 Q1 | Energy V | 0.63446 | [0.52116, 0.70907] |
| T2 Q2 / T2 Q1 | Increment Energy V | −4.11267 | [−4.42491, −3.82211] |
| T2 Q2 / T2 Q1 | Spatial VS | 0.55218 | [0.48638, 0.61895] |
| T2 Q2 / T2 Q1 | Tail Brier | 0.32078 | [−0.04168, 0.59211] |
| T2 Q2 / RootAR | Energy V | 4.21922 | [2.61675, 5.99245] |
| T2 Q0 / CSDI Q0 | Energy V | 5.65379 | [−2.14618, 16.77225] |
| T2 Q0 / TimeGrad Q0 | Energy V | 56.16823 | [52.12155, 59.70344] |
| T2 Q0 / D3U Q0 | Energy V | −2.17201 | [−4.46175, 0.12839] |
| T2 Q1 / CSDI Q1 | Energy V | 1.78257 | [−2.67936, 9.16792] |
| T2 Q2 / CSDI Q2 | Energy V | 2.04023 | [−2.41941, 9.51708] |
| T2 Q1 / TimeGrad Q1 | Energy V | 34.72032 | [29.46051, 39.99705] |
| T2 Q2 / TimeGrad Q2 | Energy V | 34.30593 | [29.20995, 39.56780] |
| T2 Q1 / D3U Q1 | Energy V | −1.31320 | [−2.97353, 0.85513] |
| T2 Q2 / D3U Q2 | Energy V | −0.83353 | [−2.43333, 1.32865] |

among 25,165,824 values; no clipping or repair is applied. Positive support does not ensure copula, rank or tail preservation. The original sampler retains its fixed $[0.05, 0.95]$ latent-quantile range. This result tests one compact, bag-trained variant, not optimization saturation.

**Development evidence boundary.** The results assign different contributions to proposal quality, generic calibration and protected control. Controllability alone does not establish optimal generation. Scoring-rule training itself is prior art (Pacchiardi et al., 2024); our contribution is the constrained attribution and protection framework, not ownership of the third-party backbones. All four matched comparisons are now complete. D3U has the lowest Energy mean, RootAR the lowest increment Energy, CSDI Q2 the lowest spatial VS, and TACTiS-2 Q2 the lowest tail Brier in the current thirteen-configuration panel; these are descriptive property leaders, not independent superiority claims. No full stack dominates all properties. A subsequent, separately frozen 2022 selection and locked 2023 confirmation are reported in Appendix M. They do not relabel these 2020 results or retrofit its statistics. MERRA-2 2021 is historical development. The separately locked 2024 follow-up appears in Appendix N; 2025 was still sealed at that study's closeout and was subsequently consumed only by Appendix O. Neither this development table nor the later two-configuration confirmation establishes fully optimized baseline coverage or field-wide SOTA.

## M LOCKED MERRA-2 TIME-HOLDOUT CONFIRMATION

**Selection and estimand.** This protocol tests whether the selected protected correction improves its calibrated input on a new time holdout. We froze it after the four-backbone 2020 development study, before opening the new selection and confirmation values. This is not study-inception preregistration. The 2018 backbone fits, normalization and reference bank, 2019 adapters, terminal checkpoints and sampling semantics remain unchanged. Selection S uses 192 common 2022 origins and all thirteen configurations in Appendix L: NativeRootAR and four Q0/Q1/Q2 stacks. Its 960 base-origin evaluations yield 5,568 replicate records. Complete S selects $c^* =$ CSDI Q2 and $u^* = \mathrm{ownQ1}(c^*) =$ CSDI Q1. Their S Energy means are 0.465572185 and 0.467244940. The 0.358% improvement is selection evidence; q99 error also increases by 18.094% in S. Neither result changes the lock.

V then evaluates only these two configurations on 192 common 2023 origins, with $K = 128$, $H = 16$, $L = 96$ and $M = 64$. Three fixed adapter replicates are averaged within each origin, producing 1,152 records but only 192 outer units. The full-quality contrast against $u^*$ and the control-increment contrast against its own Q1 coincide. We retain all twelve named slots and their

original error allocation; we do not count this pair as two independent replications. The shared Q0 sampling dependency is not a third evaluated V configuration.

**Frozen inference.** For each question and each of six lower-is-better metrics, the estimand is $\Delta = \mathrm{mean}(s_{\mathrm{Q1}} - s_{\mathrm{Q2}})$. We use the same 16,384 circular temporal-block bootstrap draws for all twelve contrasts, block length 16 origins and seed 2609124301. Bonferroni12 at family $\alpha = 0.05$ gives 99.583333% conditional intervals for absolute differences. Relative reductions are descriptive ratios of means, not percentage intervals. The zero harm margin and directional rules were fixed in advance: a positive lower bound supports improvement; a negative upper bound supports degradation; otherwise the contrast is unresolved. The month-stratified origin calendar is not a month-stratified bootstrap. The intervals depend on temporal-block assumptions and fixed fitted models, without a distribution-free or finite-sample coverage guarantee.

Table 28: All six unique V contrasts, each retained under both registered questions. $\Delta$ and intervals are in absolute score units; positive is better. Relative reductions are descriptive. The q99 error is a diagnostic, not a proper scoring rule. Under the frozen rule, its degradation prevents a profile-wide pass.

| Metric | $\Delta$ | Conditional interval | Reduction, % | Decision |
|---|---|---|---|---|
| Energy V | 0.002163462 | [0.000799766, 0.003715852] | 0.425708 | Improvement |
| Marginal CRPS V | 0.000598757 | [0.000204307, 0.001024683] | 0.195374 | Improvement |
| Increment Energy V | 0.001202990 | [0.000779179, 0.001648689] | 0.776918 | Improvement |
| Spatial variogram | 0.000212346 | [−0.000137144, 0.000591963] | 0.168412 | Unresolved |
| Tail Brier | 0.000008064 | [−0.000015949, 0.000038441] | 0.114899 | Unresolved |
| q99 amplitude error | −0.303719980 | [−0.337757894, −0.280539551] | −16.733334 | Degradation |

**Confirmed benefit and non-compensating harm.** Table 28 supports improvements in whole-path Energy, marginal CRPS and increment Energy. Spatial and coexceedance scores remain unresolved; this is not equivalence or evidence of harmlessness. The 16.733% increase in q99 amplitude error rules out the strict profile claim. The two named Energy questions are supported, but not all safeguards. Across the six unique results there are three improvements, two unresolved contrasts and one degradation; the twelve named slots duplicate these counts. This is useful property-specific evidence, not a qualification-wide pass.

Table 29: All thirteen V metrics. Entries average three replicates within each origin before averaging the 192 origins. Raw minimum and maximum entries are means of per-origin extrema, not global extrema. Diversity diagnostics have no assumed monotone preference.

| Metric | CSDI Q1 | CSDI Q2 |
|---|---|---|
| Energy V | 0.508202655 | 0.506039194 |
| Marginal CRPS V | 0.306468061 | 0.305869303 |
| Increment Energy V | 0.154841344 | 0.153638354 |
| Spatial variogram | 0.126087442 | 0.125875096 |
| Tail coexceedance Brier | 0.007018093 | 0.007010029 |
| q99 amplitude error | 1.815059561 | 2.118779541 |
| Distinct members | 64 | 64 |
| Effective rank | 46.593426513 | 46.589114505 |
| Pairwise RMS spread | 0.843878889 | 0.955364893 |
| Nearest-reference RMS | 1.174705326 | 1.209607636 |
| Reference copy matches | 0 | 0 |
| Mean per-origin raw minimum | 0.006867877 | 0.006393206 |
| Mean per-origin raw maximum | 6.834643003 | 7.098860286 |

**Temporal and sample diagnostics.** Energy, CRPS and increment Energy improve descriptively in all four quarters; q99 error worsens in all four. Spatial VS worsens in Q4, while tail Brier worsens in Q1 and Q4. No additional quarterly tests are conducted. Table 30 retains all quarterly and registered lag diagnostics. Lag counts origins, not hours. In particular, q99 differences have lag-1 correlation 0.4281. Two lag diagnostics cannot validate complete dependence coverage by block16. The 64

distinct members and zero fixed-bank matches describe this reference set, not privacy or absence of memorization. Spread increases while effective rank is nearly unchanged. Their coexistence with tail harm does not identify its causal mechanism.

Table 30: Descriptive quarter-wise relative reductions (percent) and correlations of the absolute origin-level paired difference sequence. No quarter-wise inference or post-hoc block selection is added.

| Metric | Q1 | Q2 | Q3 | Q4 | Lag 1 | Lag 16 |
|---|---|---|---|---|---|---|
| Energy V | 0.362120 | 0.301392 | 0.767280 | 0.172395 | 0.167835 | 0.023038 |
| CRPS V | 0.174855 | 0.068202 | 0.344543 | 0.167222 | 0.124168 | 0.089329 |
| Increment EV | 0.640339 | 0.641787 | 1.033615 | 0.736508 | 0.200119 | −0.003243 |
| Spatial VS | 0.061750 | 0.144763 | 0.451376 | −0.061231 | 0.146646 | −0.007986 |
| Tail Brier | −0.017937 | 0.414820 | 0.266745 | −0.343331 | 0.022978 | 0.098856 |
| q99 error | −18.467862 | −17.019359 | −14.131566 | −18.391456 | 0.428100 | 0.069139 |

**Execution and scientific boundary.** V completes once in 6,203.51 seconds with no operational recovery or scientific retry. Reconstructing the S lock and reducing all 192 contiguous V commits reproduces the stored summaries. Runtime output-validation flags and required packet presence are complete; this closeout does not independently re-certify every saved tensor. All thirteen reported metrics are finite. In that older execution, the per-run 2021, 2024–2025, new-training and scientific-retry flags were false. These flags do not establish that 2021 was never used in historical development. The later authorized 2024 study is reported separately in Appendix N.

The selected Q2-over-Q1 scoring benefit transfers to a time holdout within MERRA-2. The comparison does not rank all thirteen configurations on V or establish cross-family replication, practical importance or optimization saturation. Mean, pair-difference and support protection do not entail tail fidelity. Retaining these gains with a tail-aware correction requires a new study with a separately frozen design and fresh validation. Retuning on this consumed V panel cannot provide that evidence. The later event-preserving study uses a separate locked identity and does not replace this negative result.

## N Event-preserving multiscale correction on the locked 2024 panel

**Preserving events after the earlier tail failure.** The 2023 result showed that geometric protection alone did not ensure tail fidelity. We subsequently developed an event-preserving correction. For each adjacent channel pair, it redistributes a common shift across members. The feasible set preserves the Q1 ensemble mean, within-pair differences, support floor and every member's registered exceedance indicator. A multiscale-path controller conditions this action on eight existing slack features and nine summaries of Q1 path shape over 4, 8 and 16 steps. The matched multiscale-history arm instead uses 24, 48 and 96 observed steps. Both are 17–16–1 networks with 305 parameters, identical initialization within each of three replicates, and twelve fixed training decisions per terminal. They use the existing event projection and 2019 blockwise objectives. Parameter and budget matching does not equalize feature distributions, effective action or total optimization opportunity.

Candidate selection used adaptive 2020 development. The earlier 2023 outcome also informed the redesign. Before accessing 2024, we locked the full-strength multiscale-path candidate, all three terminals, 21 configurations and 25 named contrasts. This is post-development pre-specification, not study-inception preregistration. The 2021 split belongs to historical development, 2022 to the earlier selection study, and 2023 to the earlier confirmation study. None is relabeled as fresh evidence for the new candidate. That execution read only the exact 2024 annual partition and fixed 2018 scoring resources; 2025 was then unopened. Its later one-shot use is reported separately in Appendix O.

**Complete comparison and inference.** We evaluate 192 common 2024 origins with $K = 128$, $L = 96$, $H = 16$ and $M = 64$. Each of twelve months supplies sixteen origins, with a 112-hour year-start embargo and targets confined to their month. NativeRootAR, Root-CSDI, Root-TimeGrad, Root-D3U and Root-TACTiS-2 run serially. The original thirteen configurations remain

present, together with seven additional CSDI directions and TimeGrad path-tail. This yields 960 base-origin evaluations and 10,176 replicate records. Each base-origin shares one Q0 proposal across its adapters. The sampler namespace is 2609137101, combined with base code and the full-panel target-start index. No fit, checkpoint selection, clipping, resampling or scientific retry is added.

Replicates are averaged within each origin; the temporal sample size is 192, not 576. The estimand is mean reference score minus candidate score. All 25 contrasts share 65,536 circular-block bootstrap draws, block length 16 origins and seed 2609137301. Bonferroni25 at family level 0.05 yields 99.8% two-sided conditional intervals for absolute differences. Positive lower bounds support improvement; negative upper bounds support degradation; other intervals are unresolved, not equivalence. Relative reductions are descriptive ratios of means. The intervals condition on the fixed fitted models and temporal-block approximation. A single annual panel and seasonal dependence preclude a distribution-free coverage claim.

Table 31: All 25 locked 2024 contrasts. A: candidate versus own Q1; B: versus old event-cell; C: versus matched history; D: versus strong Q1-anchored correction. Positive $\Delta$ is better. I, D and U denote supported improvement, supported degradation and unresolved. Intervals are 99.8% conditional absolute-difference intervals. Pooled q99 error is a diagnostic, not a proper scoring rule. For A–C, Brier is checked by memberwise event identity, not a statistical test.

| Group | Metric | $\Delta$ | Conditional interval | Decision |
|---|---|---|---|---|
| A | Energy V | 0.000495952 | [0.000269918, 0.000804207] | I |
| A | CRPS V | 0.000259021 | [0.000118723, 0.000426575] | I |
| A | Increment EV | 0.000109020 | [0.000072365, 0.000147093] | I |
| A | Spatial VS | 0.000247311 | [0.000131653, 0.000379281] | I |
| A | Pooled q99 error | 0.016348170 | [0.011273929, 0.022052132] | I |
| A | Path-q99 CRPS V | 0.000515042 | [0.000180259, 0.000897820] | I |
| B | Energy V | 0.000099935 | [0.000031113, 0.000221785] | I |
| B | CRPS V | 0.000028343 | [−0.000065202, 0.000115201] | U |
| B | Increment EV | 0.000038009 | [0.000020115, 0.000057094] | I |
| B | Spatial VS | −0.000003935 | [−0.000038787, 0.000030077] | U |
| B | Pooled q99 error | 0.004098166 | [−0.000420632, 0.007481771] | U |
| B | Path-q99 CRPS V | 0.000127438 | [−0.000138851, 0.000383495] | U |
| C | Energy V | 0.000143617 | [0.000058385, 0.000274364] | I |
| C | CRPS V | 0.000048727 | [−0.000055754, 0.000142942] | U |
| C | Increment EV | 0.000047308 | [0.000025774, 0.000066024] | I |
| C | Spatial VS | 0.000001278 | [−0.000032471, 0.000034168] | U |
| C | Pooled q99 error | 0.003664008 | [−0.000528383, 0.006774832] | U |
| C | Path-q99 CRPS V | 0.000127925 | [−0.000167524, 0.000409159] | U |
| D | Energy V | −0.000153219 | [−0.000242616, −0.000051385] | D |
| D | CRPS V | −0.000042177 | [−0.000208789, 0.000064603] | U |
| D | Increment EV | −0.000059174 | [−0.000093993, −0.000025623] | D |
| D | Spatial VS | −0.000112492 | [−0.000228558, −0.000001457] | D |
| D | Tail Brier | −0.000015804 | [−0.000038859, −0.000000830] | D |
| D | Pooled q99 error | −0.007560418 | [−0.022141584, 0.002250426] | U |
| D | Path-q99 CRPS V | 0.000176503 | [−0.000782124, 0.000716934] | U |

**A supported but small quality increment.** Relative to its own Q1, the locked candidate reduces Energy V by 0.097307%, marginal CRPS V by 0.085176%, increment Energy V by 0.068603%, spatial VS by 0.193952%, pooled q99 error by 0.879100%, and path-q99 CRPS V by 0.110471%. All six absolute-difference intervals have positive lower bounds. Hard coexceedance Brier is identical memberwise through the event construction. Each of four quarterly means and each of twelve quarter-by-replicate cells also improves all six variable metrics descriptively. These are not twelve independent confirmations, nor additional quarterly hypothesis tests. The observed transfer closes the earlier gap between adaptive development and a locked temporal test. It does not establish a large practical effect.

**Strong comparisons remain unfavorable.** Against old event-cell and the matched history arm, only Energy and increment Energy have supported improvements; the other four contrasts in each group are unresolved. Relative to strong Q1-anchored correction, the candidate has supported degradations in Energy, increment Energy, spatial VS and tail Brier. The corresponding relative changes are $-0.030100\%$, $-0.037276\%$, $-0.088471\%$ and $-0.303672\%$. Its apparent path-q99 advantage over that reference remains unresolved. Across the 25 contrasts, ten support improvement, four support degradation and eleven are unresolved. These counts do not replace the four separate scientific questions.

The complete absolute panel below describes the broader comparison. It is neither a second selection stage nor an all-roster simultaneous test. The candidate has the eighth-lowest Energy mean, 0.509182291; the lowest belongs to D3U original Q2, 0.502827453. It is not first on any of the seven displayed metrics. Root-law variants are disclosed adaptations of published methods, not unmodified released laws, compute-matched tuning or optimization-saturated baselines. Toto's earlier ineligibility is not a victory against an eligible comparator.

Table 32: Complete locked 2024 absolute-quality panel, part 1. Lower is better. All three adapter replicates are retained. No winner is selected on this panel.

| Configuration | Energy V | CRPS V | Increment EV |
|---|---|---|---|
| NativeRootAR | 0.522977343 | 0.323917978 | 0.158423820 |
| CSDI Q0 | 0.536554482 | 0.308049124 | 0.159493494 |
| CSDI Q1 | 0.509678244 | 0.304101218 | 0.158914585 |
| TimeGrad Q0 | 1.138895548 | 0.618498899 | 0.280740569 |
| TimeGrad Q1 | 0.751118003 | 0.422182691 | 0.261294850 |
| D3U Q0 | 0.506174183 | 0.306234849 | 0.158986725 |
| D3U Q1 | 0.503508509 | 0.304842484 | 0.159143571 |
| TACTiS-2 Q0 | 0.513491986 | 0.308350249 | 0.270814784 |
| TACTiS-2 Q1 | 0.513294658 | 0.308759447 | 0.263344705 |
| CSDI original Q2 | 0.507650722 | 0.303458660 | 0.157685413 |
| TimeGrad original Q2 | 0.739666201 | 0.420891148 | 0.268970247 |
| D3U original Q2 | 0.502827453 | 0.303924656 | 0.158857270 |
| TACTiS-2 original Q2 | 0.510031529 | 0.307328346 | 0.273677826 |
| CSDI path-tail | 0.507737916 | 0.303343416 | 0.158016953 |
| CSDI Q1-anchored | 0.509029073 | 0.303800019 | 0.158746391 |
| CSDI block-anchored | 0.509666722 | 0.304087741 | 0.158912252 |
| CSDI event projection | 0.509057835 | 0.303803263 | 0.158737522 |
| CSDI event-cell | 0.509282226 | 0.303870540 | 0.158843574 |
| CSDI multiscale-path | 0.509182291 | 0.303842196 | 0.158805565 |
| CSDI multiscale-history | 0.509325908 | 0.303890924 | 0.158852873 |
| TimeGrad path-tail | 0.748782431 | 0.422855397 | 0.253001413 |

**Variation, action and limits of protection.** The candidate and Q1 both retain 64 distinct members and zero matches to the fixed reference bank. Effective rank rises from 45.2326 to 45.4665 and pairwise RMS spread from 0.858818 to 0.872362. This is not a privacy or non-memorization guarantee, and more spread is not intrinsically better. Path-q99 spread decreases from 0.821657 to 0.818820; mean path-q99 signed error rises from 0.017266 to 0.019719. Full 23-field diagnostics remain in the machine-readable summary.

Pooled action RMS, defined as the square root of mean squared action over all candidate origin–replicate records, is 0.110333 in the registered units. Old event-cell, Q1-anchored and event-projection values are 0.069844, 0.116214 and 0.113642. These describe action size, not exhausted headroom or an equal-action causal comparison. Notably, the old event projection shares the candidate's event constraint yet has better mean scores on five of the six variable metrics; the candidate has better path-q99 CRPS V. These results do not identify event protection alone as the cause of the remaining quality gap. The next development question concerns how objective allocation and effective action influence that trade-off within the same feasible set.

**Verification and claim boundary.** The run completes once in 12,797.53 seconds, with zero operational recovery or scientific retry. All 960 base-origin commits reproduce the stored aggregate after

Table 33: Complete locked 2024 absolute-quality panel, part 2. Lower is better. All three adapter replicates are retained. No winner is selected on this panel.

| Configuration | Spatial VS | Tail Brier | Pooled q99 error | Path-q99 CRPS V |
|---|---|---|---|---|
| NativeRootAR | 0.137457165 | 0.005564998 | 2.062193930 | 0.465356985 |
| CSDI Q0 | 0.132680059 | 0.005395110 | 1.854907043 | 0.492422843 |
| CSDI Q1 | 0.127511400 | 0.005220090 | 1.859647776 | 0.466224909 |
| TimeGrad Q0 | 0.317699732 | 0.007681320 | 4.534494590 | 1.228868656 |
| TimeGrad Q1 | 0.200952183 | 0.006601393 | 3.195739331 | 0.805638867 |
| D3U Q0 | 0.129923096 | 0.005184567 | 2.356578228 | 0.449459494 |
| D3U Q1 | 0.129600019 | 0.005154575 | 2.403674524 | 0.452004720 |
| TACTiS-2 Q0 | 0.131666624 | 0.005193337 | 1.259344804 | 0.676383517 |
| TACTiS-2 Q1 | 0.131009789 | 0.005177715 | 1.159999134 | 0.629444864 |
| CSDI original Q2 | 0.127187726 | 0.005228511 | 2.158144915 | 0.466409391 |
| TimeGrad original Q2 | 0.198710084 | 0.006553780 | 3.251499931 | 0.830901374 |
| D3U original Q2 | 0.128825183 | 0.005160504 | 2.241251937 | 0.450175139 |
| TACTiS-2 original Q2 | 0.130356387 | 0.005160493 | 1.254399169 | 0.641770322 |
| CSDI path-tail | 0.127163820 | 0.005227823 | 2.162217524 | 0.464501097 |
| CSDI Q1-anchored | 0.127151597 | 0.005204286 | 1.835739189 | 0.465886370 |
| CSDI block-anchored | 0.127498613 | 0.005219479 | 1.859372479 | 0.466209778 |
| CSDI event projection | 0.127157821 | 0.005220090 | 1.835346790 | 0.465916197 |
| CSDI event-cell | 0.127260155 | 0.005220090 | 1.847397772 | 0.465837305 |
| CSDI multiscale-path | 0.127264089 | 0.005220090 | 1.843299606 | 0.465709867 |
| CSDI multiscale-history | 0.127265367 | 0.005220090 | 1.846963614 | 0.465837793 |
| TimeGrad path-tail | 0.200398020 | 0.006588844 | 3.091700110 | 0.763226731 |

Table 34: Descriptive quarterly relative reductions against the candidate's own Q1, in percent. All three replicates are averaged within origins. No quarter is removed and no quarter-specific interval is added.

| Metric | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| Energy V | 0.091112 | 0.113341 | 0.107183 | 0.071798 |
| CRPS V | 0.105097 | 0.063035 | 0.123176 | 0.040139 |
| Increment EV | 0.072188 | 0.093767 | 0.055323 | 0.056488 |
| Spatial VS | 0.161992 | 0.184724 | 0.249166 | 0.158322 |
| Tail Brier | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Pooled q99 error | 0.626038 | 1.167275 | 0.665803 | 1.208188 |
| Path-q99 CRPS V | 0.066567 | 0.144558 | 0.151112 | 0.043431 |

JSON sequence normalization. Independent direct averaging differs by at most $2.85 \times 10^{-14}$; explicit origin expansion of the same bootstrap draws reproduces the 25 intervals within $4.86 \times 10^{-17}$. Re-scoring first, middle and last origins across five bases reproduces 159 records and 3,657 metric entries exactly. The same spot check verifies 36 event-preserving path sets with no changed indicator. This is a bounded check of saved samples, not an independent second experiment.

The supported claim is a small, event-preserving, six-metric improvement over an already calibrated Q1 on one locked year. It neither replaces the old 2023 harm result nor establishes superiority over all strong comparators, cross-family replication of the new controller, or general $K > 2$ dependency protection. The paired-slack invariants and the thin KRPC theorem retain their separate scopes. The 2024 panel is now consumed; further model development must not use it as a fresh confirmation. The 2025 boundary was closed at this closeout, before the separately frozen contextual-gate study.

# O FRESH 2025 CONTEXTUAL FEASIBLE-CONE CONFIRMATION

After completing all 2020 gate development, we froze a target-free contextual gate and opened 2025 exactly once. The gate chooses between protected corrections of the same generated ensemble. The study used 192 common origins, $K = 128$, $L = 96$, $H = 16$, $M = 64$, and the unchanged 21-arm 2024 roster plus the new candidate. For each origin the gate made one past-measurable choice between the mature multiscale expert and the new feasible-cone expert, shared across three replicates. It did not observe the current target, select a replicate, refit a model, or alter a threshold.

The gate chose the feasible-cone expert on 141 origins and the mature expert on 51. Relative to the mature expert, mean Energy V decreased from 0.4600244293 to 0.4597483352, a 0.0600% reduction. The preregistered one-sided 95% circular-block-bootstrap lower bound for the absolute reduction was 0.0001986. All five secondary ratios were below one; the largest quarterly Energy ratio was 0.9996613; the largest of six ratios to Q1 was 0.9985074. Tail Brier was bitwise equal, and all event, support, diversity, non-copying and controller certificates passed. The internal conjunction therefore passed.

The broader generator comparison limits that success. The candidate ranked third among 22 configurations. Its Energy was 0.3344% above CSDI affine-structure and 0.3302% above CSDI affine-path-tail. Sixteen of 21 paired comparisons had a positive Bonferroni-adjusted lower bound, but the two lower-mean affine arms and three statistically unresolved D3U comparisons prevent the frozen all-roster claim.

Among the five exact-event configurations the candidate had the lowest descriptive mean and each corresponding precomputed Bonferroni comparison was positive; because that subgroup was not a separate preregistered SOTA estimand, this is only a historical within-subgroup lead, not a replacement SOTA declaration. The expanded post-hoc controls below also refute a general event-constrained Energy lead. The original result establishes a fresh, non-compensating improvement of a protected controller. It also falsifies whole-roster superiority.

## O.1 FROZEN GATE AND EXPERT SPECIFICATION

The gate is a fixed ridge regressor, not an additional generative backbone. Let $F$ denote the feasible-cone expert and $R$ the mature multiscale-path expert. Both operate on the same calibrated CSDI Q1 ensemble for each replicate; both use full strength $\gamma = 1$. For origin $i$ and zero-based calendar month $m_i$, define $\phi_i = 2\pi(m_i + 1/2)/12$. A bar below denotes the equal mean across the three adapter replicates, not a choice among them. The eight-dimensional gate input is

$$f_i = \big(\sin\phi_i, \cos\phi_i, \overline{A}_F, \overline{A}_R, \overline{\Delta V}_F - \overline{\Delta V}_R, \overline{W}_F - \overline{W}_R, \overline{E}_F - \overline{E}_R, \overline{T}_F - \overline{T}_R\big). \tag{29}$$

For an expert output $Z$ and its Q1 input $X$, with the fixed physical unit $u$, $A$ is the mean of $((Z - X)/u)^2$ over all member, horizon and channel entries. $\Delta V$ is the difference between their member-centered variances, likewise averaged over entries in units $u^2$. To define $W$ and $E$, flatten each path after the fixed 2018 scoring standardization into $b_a \in \mathbb{R}^D$, $D = HK$. Then $W = [(M^2 D)^{-1} \sum_{a,b} \|b_a - b_b\|^2]^{1/2}$, including diagonal zeros. $E = \exp(-\sum_j p_j \log p_j)$ uses eigenvalues of the centered Gram matrix $BB^\top/D$, retaining those greater than $10^{-12}$ times its largest nonnegative eigenvalue and normalizing the retained values to $p_j$; it is zero when no positive mass remains. Finally, let $q_{a,k}$ be the linear 0.99 quantile of physical path $a$ over its $H$ steps. $T$ is

the channel mean of the population standard deviation of $q_{a,k}$ across members, divided by $u$. None of these features uses the verifying target.

On the 192 already-consumed 2020 development origins, the training label is $y_i = \frac{1}{3}\sum_{r=0}^{2}[S_{\mathrm{V}}(Z_{ir}^{R}, Y_i) - S_{\mathrm{V}}(Z_{ir}^{F}, Y_i)]$. Each of twelve leave-one-month-out fits trains on 176 origins and tests on the remaining sixteen. Feature means $\mu_j$ and population standard deviations $\sigma_j$ are calculated only on that training fold; a zero standard deviation is replaced by one. With $z_{ij} = (f_{ij} - \mu_j)/\sigma_j$, the fit solves

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0,\beta} \sum_{i\in\mathcal{I}_{\mathrm{train}}} (y_i - \beta_0 - z_i^{\top}\beta)^2 + \|\beta\|_2^2. \tag{30}$$

The loss is a sum, not a mean; the ridge coefficient is one and the intercept is unpenalized. A strictly positive prediction selects $F$; zero or negative selects $R$. After the twelve-fold development check, the identical rule was fit once on all 192 origins, saving its nine coefficients and eight-element center and scale vectors. Features, regularization, threshold and expert strength were not searched. Cross-fitting does not restore freshness after adaptive development. The final coefficients, not the cross-fit fold models, were frozen for 2025; there was no 2025 fitting or target-dependent switching.

At inference, both expert outputs for all three replicates are generated before constructing $f_i$. The one decision selects each chosen expert's complete ensemble for all replicates. It neither averages experts nor selects members. Here "target-free" means a measurable function of the available history, calendar, frozen resources and generated random tape; generated future paths are not observed future measurements. Targets are used only after this choice for verification.

**The two experts and their protected action.** Both experts use the same 305-parameter 17–16–1 tanh network and the same event-cell projection. The difference is the fitted parameters and training path, not a new inference-time constraint solver. For adjacent channel pair $p$ at member $a$ and time $h$, write $s = \min(X_{ah,2p}, X_{ah,2p+1})$, $\ell = \log(1+s)$, and $v = (X_{ah,2p} + X_{ah,2p+1})/2$, in physical-unit coordinates. The original eight inputs are $\ell$, its member-centered value, $\log(1 + |X_{ah,2p} - X_{ah,2p+1}|)$, its time-centered value, $(v - \overline{v}_{\mathrm{members}})/(1 + \overline{v}_{\mathrm{members}})$, the pair-mean of the member-centered $\ell$, the log-one-plus last observed pair level, and $h/(H-1)$ for zero-based $h$. Nine extra inputs summarize $\log(1+v)$ over the trailing 4, 8 and 16 generated steps: for each window, tanh of its mean minus last value, population standard deviation divided by one plus itself, and tanh of its least-squares slope against equally spaced times in $[-1, 1]$. These nine values are broadcast over $h$.

The network output is bounded as $a_\theta = 2\tanh(\mathrm{MLP}_\theta(f))$. For each time and pair, form the proposal

$$v_a^{\mathrm{prop}} = s_a \exp(a_{\theta,a}) \frac{\sum_b s_b}{\sum_b s_b \exp(a_{\theta,b})} - s_a;$$

when all slack is zero the proposal is zero. The common pair shift is the Euclidean projection of this vector onto $\{d : \sum_a d_a = 0,\ L_a \le d_a \le U_a\}$. Before event restrictions, $L_a = -0.99 s_a$ and $U_a = 0.99\sum_{b\neq a} s_b$. For each of the two channels, let $e = X - \tau$, using the fixed fit-derived threshold in the same units. If $e > 0$, intersect the lower bound with $d \ge -0.9e$; otherwise intersect the upper bound with $d \le -0.9e$. The projected solution has form $d_a = \min(U_a, \max(L_a, v_a^{\mathrm{prop}} - \lambda))$ with zero sum. The frozen implementation uses 80 bisection iterations, one active-face affine refinement and an exact fixed-point branch for already feasible proposals. It supports first-order proposal gradients on a fixed face, not derivatives through the fixed Q1 or threshold. The output adds $d_a$ to both channels. Thus the specified mean, pair difference, support floor and every registered hard-event indicator are protected, subject to the recorded numerical certificates. Neither pooled-q99 amplitude nor all cross-channel dependence is an exact invariant; these remain scored properties. This is not the thin KRPC operator and does not inherit its full complement or endpoint theorem.

**Training path and budget.** The mature expert was fitted on 256 fixed 2019 origins in four chronological blocks of 64, with at most twelve accepted updates per replicate. It used the seed $26091024 + r$, a zero output head and the same six real objectives: Energy V, marginal CRPS V, increment Energy V, spatial variogram, path-q99 CRPS V and pooled-q99 absolute error. The new expert instead transplanted the fitted eight-input Q1-anchored entropy policy into the first eight columns of the 17-input network, setting the extra columns to zero. This preserves the initial score function, not necessarily its output before event projection. It allowed at most 24 restoration updates

followed by twelve quality updates per replicate, beyond upstream fitting; these are not equal-total-compute runs.

Restoration compares thirty cells (six overall objectives and six in each block) with original Q1. Violated cells must strictly decrease and satisfy Armijo with coefficient $10^{-4}$; unviolated cells must remain no worse than Q1. A common descent direction is obtained from the minimum-norm convex combination of selected unit-normalized gradients and then negated and unit-normalized. The solver is SLSQP with tolerance $10^{-15}$ and at most 2,000 iterations, followed by one active-face polish and explicit KKT/slope checks. In at most thirty guard rounds, an unviolated cell is added when its positive directional slope times the smallest step exceeds its Q1 slack. This is the parameter-space "feasible cone", not the event projection. After restoration, the quality phase requires strict Armijo decrease of all six metrics in each block and overall, no increase over original Q1, and unchanged hard-event Brier. Both phases try the same eight ordered step lengths $0.05\,2^{-j}$, $j = 0, \ldots, 7$, stopping on an uncertified direction, failed line search or budget exhaustion. A stop does not prove global infeasibility or optimality. All three terminals are retained.

The executable definitions are in `merra2_contextual_expert_gate.py`, `merra2_multiscale_conditioning_fit.py`, `event_cell_projection.py`, and `merra2_feasible_cone_formal_fit.py` with its direction/acceptance helpers. The 2025 backend constructs target-free features before scoring. This specification documents frozen behavior; it does not replace missing weights, data access instructions or a complete executable reproduction package.

### O.2 Post-hoc matched event projection of the strongest affine controls

The two stronger affine controls might benefit from being free to change hard-event identities. We tested this explanation on the already exposed 2025 panel after freezing two additional controls and their evaluation. This is a post-hoc comparison, not a new confirmation or a revision of the original 22-arm decision. For each original Q1 ensemble, we took each affine control's cached displacement, averaged the displacement within each adjacent channel pair, and projected it with the same zero-sum, support and event-cell constraints described above. We retained the original full action scale, thresholds, $0.9$ event margin and numerical solver; we did not search strengths or train another model.

The experiment reused all 192 origins and three technical replicates, with $M = 64$, $H = 16$ and $K = 128$. All 1,152 projected ensembles were persisted before scoring against targets, with no new generative random draws. Independent checks verified the projection and KKT conditions, finite outputs, 64 distinct members, non-copying, mean and pair certificates, and zero hard-event flips. All 23 original score fields were replayed, including 576 original Q1 score checks; Energy and marginal CRPS additionally received independent explicit arithmetic checks. Replicates were averaged within origin before averaging origins. No new confidence interval or significance claim is attached to this adaptive comparison.

| Configuration | Energy | Marginal CRPS | Pooled-q99 error | Path-q99 CRPS |
|---|---|---|---|---|
| Q1 | 0.460449 | 0.280442 | 1.594254 | 0.420549 |
| Contextual gate | 0.459748 | 0.279940 | 1.575111 | 0.419921 |
| Affine structure | 0.458216 | 0.279530 | 1.870124 | 0.420962 |
| + event projection | 0.458267 | 0.279504 | 1.840371 | 0.421036 |
| Affine path-tail | 0.458235 | 0.279316 | 1.865677 | 0.418573 |
| + event projection | 0.458299 | 0.279290 | 1.833503 | 0.418654 |

Table 35: Post-hoc matched event constraints on the exposed 2025 panel. All entries are errors or scores for which lower is better. The two q99 columns use physical units. The original one-shot decision is unchanged.

Both projected affine controls still have lower Energy than our gate, by 0.3222% and 0.3153%, respectively, with the same direction in all twelve monthly and all four quarterly means. Their increment Energy improves by 0.7174% and 0.4979%, and spatial variogram by 0.0604% and 0.0907%. Their hard-event Brier is exactly the same as Q1 and our gate. Thus event preservation alone cannot

explain the original Energy gap. Adding projection increases each affine control's Energy by only 0.0112% and 0.0139%.

The tradeoff remains material: their pooled-q99 absolute errors exceed our gate's by 16.8407% and 16.4047%. Path-q99 CRPS is worse by 0.2654% for projected structure but better by 0.3018% for projected path-tail. Exact event membership does not imply exact q99 amplitude or its predictive accuracy. These results support a measured quality–tail-error tradeoff, not optimality, unique necessity of the learned policy, or SOTA under event constraints. Neither a favorable endpoint nor a post-hoc tolerance chosen from this table can establish those stronger claims.

### O.3 Matched strategy learning on the exposed development panel

To test whether policy complexity adds value within the same protected expert pool, we fitted constant, linear and nonlinear simplex policies on 2019, then evaluated the already exposed 2020 development panel. The common pool contains Q1, event-projected affine structure and event-projected affine path-tail. All policies start with uniform predictions and share six past-only features from 96 hours, a 2019-only feature scaler, upstream ensembles, batch tapes, loss and optimizer. Each of three technical replicates receives 400 Adam updates with batch size four. The policies have 3, 21 and 163 parameters: matched updates and data access do not imply matched FLOPs. This is adaptive development, not a fresh annual confirmation.

We retained all seven arms, including the three single experts and uniform mixing, at the original 192 origins with $M = 64$, $H = 16$, $K = 128$. All 4,032 ensembles were persisted before target scoring. An independent audit checked 576 seven-arm packets, 192 score records and 2,774 bound files, independently replaying features, policy weights, projection, paths and all 23 score fields. Explicit arithmetic for the seven core quality metrics agreed within $7.8 \times 10^{-16}$. Every ensemble retained 64 distinct members, zero reference copies and the original event, mean, pair and support certificates. Technical replicates are averaged within origin; origins receive equal weight but are not assumed temporally independent.

| Arm | Energy | Marginal CRPS | Pooled-q99 error | Path-q99 CRPS |
|---|---|---|---|---|
| Q1 | 0.527113829 | 0.313967659 | 1.741106534 | 0.441446792 |
| Projected structure | 0.525135490 | 0.313366488 | 2.002148527 | 0.442019549 |
| Projected path-tail | 0.525158502 | 0.313193278 | 1.997552664 | 0.439802751 |
| Uniform | 0.525494691 | 0.313245209 | 1.907234549 | 0.440493356 |
| Constant | 0.525257453 | 0.313229866 | 1.958543982 | 0.440489891 |
| Linear | 0.525233662 | 0.313242299 | 1.961063399 | 0.440488725 |
| Nonlinear | 0.525141799 | 0.313324544 | 1.996759672 | 0.441130240 |

Table 36: All seven matched-strategy development arms. Lower is better; the q99 columns use physical units. No significance or SOTA claim is attached to the rounded means.

Nonlinear reduces mean Energy by 0.022019% against constant, 0.017490% against linear, 0.067154% against uniform and 0.374118% against Q1. However, it ranks second of seven: projected structure remains better by 0.001201% relative to that expert. The small policy-complexity benefit does not establish added value over the best single expert. Nonlinear improves Energy over constant and linear in all three technical replicates, but only 7/12 and 8/12 months, respectively, and 3/4 quarters for each. Against projected structure it improves in 0/3 replicates, 5/12 months and 1/4 quarters. Against projected path-tail it improves by 0.003181% overall, with favorable directions in 2/3 replicates, 9/12 months and 2/4 quarters. The corresponding counts against uniform are 3/3, 10/12 and 3/4; against Q1 they are 3/3, 11/12 and 4/4. These overlapping summaries are descriptive, not independent replications or additional selection criteria.

The Energy gain comes with costs in other properties. Against constant and linear, respectively, nonlinear improves increment Energy by 0.030571% and 0.042194%, but worsens marginal CRPS by 0.030226% and 0.026256%, spatial variogram by 0.004385% and 0.010523%, pooled-q99 error by 1.951230% and 1.820251%, and path-q99 CRPS by 0.145372% and 0.145637%. Pooled-q99 error is 14.683371% worse than Q1, whereas hard-event Brier is identical across all seven arms. The complete development report retains all 23 fields, including spread, effective rank, support

extrema, reference-distance/copy diagnostics, and signed, overshoot, undershoot and range-based q99 diagnostics, together with all replicate, monthly and quarterly means.

Protected structure does not guarantee predictive or tail-amplitude improvement. These fixed-budget results do not establish an optimization ceiling. We retain them unchanged and report a separate, larger-budget sensitivity below; neither comparison is promoted to the consumed 2025 panel or treated as fresh confirmation.

**Additional optimization sensitivity.** Starting from each 400-update nonlinear model and its Adam state, we ran one bounded extension of 800 updates at batch size 16 on the same 256 FIT origins in 2019. Architecture, past-only features, experts, learning rate and Energy objective were unchanged. We used the terminal 1,200-update checkpoint, without epoch selection, and did not retrain the seven reference arms. Thus this evaluates additional optimization, not equal-compute efficiency or an isolated architecture effect. On FIT, Energy decreased by 0.017295% relative to the 400-update policy. Reusing the audited expert caches, we evaluated 576 additional ensembles on the same 192 exposed 2020 origins; all original protection, diversity and non-copying checks passed. Independent policy/path reconstruction and core score arithmetic agreed within $7.2 \times 10^{-15}$ and $7.8 \times 10^{-16}$, respectively.

The extended policy attained Energy 0.525091780, marginal CRPS 0.313311185, pooled-q99 error 2.002672648 and path-q99 CRPS 0.441270443. Its mean Energy was lowest among the resulting eight arms, improving by 0.009525% over the 400-update policy and 0.008324% over projected structure, the strongest fixed expert by mean Energy. The extended policy had the lowest Energy in each of the three technical replicates. Its improvement over projected structure occurred in only 4/12 months and 2/4 quarters. Against the 400-update policy, Energy improved in 8/12 months and all four quarters, whereas pooled-q99 error worsened by 0.296129% overall, in 11/12 months and every quarter. Relative to projected structure, pooled-q99 error worsened by 0.026178% and increment Energy by 0.033040%, despite a 0.169473% improvement in path-q99 CRPS. Hard-event Brier remained unchanged. The complete eight-arm report retains all 23 fields and replicate, monthly and quarterly summaries.

This small adaptive-development Energy lead shows that the 400-update result was not an optimization ceiling; it does not establish non-compensating dominance, temporal robustness, irreplaceability or field-wide SOTA.

**Tail-aware continuation: objective sensitivity, not promotion.** To distinguish more optimization from a change in objective, we continued each 1,200-update model and its complete Adam state for 400 further updates at batch size 16. The two arms share the original 256 FIT origins, three protected experts, six past-only features, architecture, optimizer settings and a persisted batch tape within each technical replicate. One minimizes Energy alone; the other minimizes $J = E + 0.1(E_0/T_0)T$, where $T$ is empirical V-form CRPS of each channel's pathwise temporal q99, averaged over channels. The fixed scales $E_0 = 0.5020155273830468$ and $T_0 = 0.43399587121942285$ are the prior 1,200-update policy's complete FIT means, not mini-batch or future-data statistics. We use linear quantile interpolation, the terminal checkpoint and no coefficient search. Scoring-rule minimization is existing methodology (Pacchiardi et al., 2024); this bounded comparison tests a particular protected policy and objective, not a new general training principle.

| Policy on 2019 FIT | Energy | Pooled-q99 error | Path-q99 CRPS |
|---|---|---|---|
| Nonlinear, 1,200 updates | 0.502015527 | 1.972542248 | 0.433995871 |
| Energy continuation, 1,600 | 0.502013038 | 1.970482755 | 0.434017126 |
| Tail-aware continuation, 1,600 | 0.502033469 | 1.962733337 | 0.432447365 |

Table 37: Bounded objective sensitivity on training data, not a new validation panel. Lower is better; q99 columns use the original physical units. Means average three technical replicates within each of 256 origins, then weight origins equally. All ten arms and 23 fields are retained in the complete report; historical references keep their original budgets.

Relative to the 1,200-update policy, the tail-aware arm reduces path-q99 CRPS by 0.356802% and pooled-q99 error by 0.497273%, but increases Energy by 0.003574%. Against the Energy-only

continuation, the respective tail reductions are 0.361682% and 0.393275%, with an Energy increase of 0.004070%. The tail gains over both references appear in all 12 monthly and four quarterly FIT aggregates, whereas Energy improves in only five months and one quarter. These are dependent descriptive summaries, not independent confirmations.

Replicate-level exceptions also remain: against 1,200 updates, replicate zero's path-q99 CRPS increases by 0.001595%; against Energy continuation, its pooled-q99 error increases by 0.066129%. Across all replicates, the tail-aware arm still trails the path-tail expert in path-q99 CRPS by 0.029245% and Q1 in pooled-q99 error by 14.255546%. Thus a favorable comparison to projected structure does not establish dominance over the complete reference set.

All six terminal models retain the original protection, 64 distinct members and zero reference copies. The independent audit verifies all 2,400 added checkpoint/commit pairs and optimizer counts, six first-step replays, and 1,536 terminal reconstructions, but does not replay every gradient. The full report computes 23 fields for all 7,680 origin–replicate–arm cells, crosschecks three core fields against saved audits, and independently checks seven metrics at 60 first/last-origin anchors; it is not a full independent reimplementation of all fields.

The predeclared FIT investment rule requires all three core means to be non-increasing versus 1,200 updates, and Energy and path-q99 CRPS to be non-increasing versus the Energy continuation, with strict improvement in the latter tail score. Both Energy conditions fail. We therefore close this candidate without further coefficient search or 2020 evaluation. The result shows objective-sensitive quality tradeoffs, not held-out superiority, an exhausted capability frontier, or SOTA.

### O.4 Tail-priority transfer with transparent secondary costs

The preceding conclusion records the original FIT decision. A separately authorized development evaluation then asked a narrower question: does the archived tail-aware policy improve its intended tail score when a small Energy increase is reported rather than treated as an automatic veto? We retained the original failed decision and fixed the new primary endpoint to path-q99 CRPS before evaluating this policy on the exposed 2020 cache. This is outside the policy's 2019 fitting sample, but not fresh confirmation. Both new arms use their original 1,600-update terminal weights; the tail coefficient remains 0.1. No backbone, policy, coefficient or action strength was retrained or selected on these 2020 results.

We reused 192 origins with $M = 64$, $H = 16$ and $K = 128$. Each origin averages three technical replicates before the equal-origin mean. The root ensemble is shared across replicates, not three independent draws. We kept the eight previous development arms, both continuations, root CSDI and two unprojected experts. Table 38 reports every arm. Energy and the remaining quality fields are secondary costs under this protocol; the original hard-event membership, ensemble means, paired-channel differences, support and non-copy checks remain required for the new policies. Unprojected references remain visible without assigning them these Q1-relative protection promises.

Against the matched Energy-only continuation, tail-aware training reduces path-q99 CRPS by 0.314535% and pooled-q99 error by 0.413691%, while Energy increases by 0.003026%. The primary reduction has the same direction in all three technical replicates, twelve monthly means and four quarterly means. These dependent summaries are descriptive, not independent confirmations. The policy still ranks third on the primary endpoint: it trails projected and unprojected path-tail by 0.016556% and 0.032540%, respectively. Its Energy is lower than those experts' by 0.011813% and 0.011297%.

A post-hoc two-objective check clarifies the trade-off without changing the primary ranking. Among all thirteen arms, Energy-only continuation, tail-aware continuation and unprojected path-tail are non-dominated in Energy and path-q99 CRPS: none is matched or improved on both by another arm with at least one strict improvement. For the ten common protected arms, projected path-tail replaces its unprojected counterpart on this frontier. This is a property of the observed finite-roster means, not a population optimum, a deployment recommendation or SOTA. We record all observed Energy-budget change points rather than choosing a favorable acceptance threshold after seeing 2020.

The estimand distinction matters. Path-q99 CRPS assesses distributions of channel-wise trajectory summaries; it does not assess the complete joint law or every meaning of tail fidelity. Root CSDI

| 2020 development arm | Path-q99 CRPS | Energy-V | Pooled-q99 error |
|---|---|---|---|
| Q1 | 0.441446792 | 0.527113829 | 1.741106534 |
| Projected structure | 0.442019549 | 0.525135490 | 2.002148527 |
| Projected path-tail | 0.439802751 | 0.525158502 | 1.997552664 |
| Uniform mixture | 0.440493356 | 0.525494691 | 1.907234549 |
| Constant policy | 0.440489891 | 0.525257453 | 1.958543982 |
| Linear policy | 0.440488725 | 0.525233662 | 1.961063399 |
| Nonlinear, 400 updates | 0.441130240 | 0.525141799 | 1.996759672 |
| Nonlinear, 1,200 updates | 0.441270443 | 0.525091780 | 2.002672648 |
| Root CSDI | 0.448293782 | 0.555356238 | 1.645914660 |
| Unprojected structure | 0.441948894 | 0.525147150 | 2.037074743 |
| Unprojected path-tail | 0.439732478 | 0.525155792 | 2.036390264 |
| Energy continuation, 1,600 | 0.441263492 | 0.525080575 | 2.001806817 |
| Tail-aware continuation, 1,600 | 0.439875565 | 0.525096464 | 1.993525525 |

Table 38: All thirteen arms on exposed 2020 development. Lower is better. Path-q99 CRPS is primary; both q99 columns use physical units. The complete 23-field report retains all technical-replicate, monthly and quarterly means. Historical policies retain their original budgets. No application cost threshold or new statistical significance claim is inferred.

has lower pooled-q99 error than the tail-aware policy, whose error is 21.119617% higher on that endpoint. The original protection checks, 64 distinct members and zero exact reference copies all pass for the new policies, but finite-reference non-copying is not a privacy guarantee. The independent audit rechecks all 576 deterministic policy packets, 192 complete score records, sources, protection certificates and equal-weight aggregates; it recomputes seven core metrics independently. The evidence supports objective-specific transfer with measured costs, not superiority over all strong experts.

# P Request learning, observability, and protected geometry

These studies separate executing a requested transformation, learning a request from history, and improving the resulting joint path distribution. Successful execution does not settle the latter questions. The evidence progresses from synthetic qualification to fit-only execution and target-informed development; these stages are not independent confirmation. Negative and unresolved outcomes remain alongside successful constructions.

## P.1 Fit-only learned operator execution at $K = 500$

This mechanism experiment tests learned execution of a specified operator. It was frozen after the USGS outer result had been consumed and cannot update the zero-request ranking. The D3U terminal law generated $M = 64$ paths for 32 early fit-only contexts; eight later contexts and 16 disjoint bounded-SPD requests per context formed a fixed operator test. A rank-4 request-conditioned native fiber and a 20,999-parameter matched arm without request geometry each received 5,000 updates. Both arms shared the same paths, analytic operator targets, batches, initialization and optimizer budget. Neither read a realized future or a quality score.

On the 128 unseen context–request cells, normalized error relative to the analytic thin-coordinate target is 0.096359 for the requested arm, compared with 1.0 for identity and 1.079478 for the matched unrequested arm; the requested/unrequested ratio is 0.089264 and the requested 95th percentile is 0.357096. Identity remains bit-exact. The largest coordinate, protected-complement and endpoint errors after the $K = 500$ lift are $1.89 \times 10^{-15}$, $6.66 \times 10^{-16}$ and $2.22 \times 10^{-15}$, without a $K \times K$ action. This establishes learned execution of an externally specified operator at high dimension. It does not establish how to choose a factual request, improve D3U's forecast law, or upgrade the scoped SOTA claim.

### P.2 A FALSIFIED ISOTROPIC FACTUAL POLICY

We next tested whether a history-based policy could choose a useful request. This was a frozen post-outer, fit-only experiment. The request family was restricted to the gauge-invariant scalar form $G = gI$. Eight annual blocks from 2006–2013 compiled annual weak labels on a 17-point reciprocal gain grid, 2014 was unused, and a fixed ridge map predicted one gain per origin from only its available history and caller-owned $q_0$ paths. Evaluation used four later fit years, 2015–2018. The D3U backbone had already been fitted on all 2000–2018 rows, so this is controller development, not an independent forecast test.

The contextual exact-KRPC arm reaches mean fair normalized Energy 0.457667, versus 0.457799 for identity $q_0$ and 0.457724 for the globally selected scalar gain. These are ratios 0.999712 and 0.999875, respectively; the contextual arm wins only 2/4 annual blocks and fails the frozen practical gate. Even a descriptive annual-block hindsight choice on the same grid reaches only a 0.0561% reduction relative to identity. The learned native executor retains 74.3% of the contextual arm's very small exact improvement and all lift certificates pass, but execution fidelity cannot turn a failed factual policy into positive evidence. We therefore reject this annual-label isotropic specification without retry. A richer future controller must introduce directional, gauge-equivariant context structure on fresh data.

### P.3 A CONSTRUCTIVE DIRECTIONAL REQUEST RESULT

Before exposing another real target, we tested whether that directional request can be represented without naming the four thin axes. We froze the two-coefficient family

$$G(c) = \exp\{\beta_0(c)I + \beta_1(c)S(c)\}\,,$$

where $S(c)$ is the unit-Frobenius trace-free part of a context covariance. The known-truth construction used $K = 500$, $r = 4$, $M = 64$, 512 training contexts and 256 disjoint evaluation contexts. Training labels were recovered from paired exact actions by least-squares gain identification and a symmetric matrix logarithm; a fixed ridge map then predicted $(\beta_0, \beta_1)$ from four context-scale invariants. Evaluation targets were excluded from that fit.

Normalized action error is $6.20 \times 10^{-11}$ for the proposed request, versus 0.898744 for a learned scalar-only arm and 1.256826 for a fixed-axis directional negative control. The proposed/scalar error ratio is $6.89 \times 10^{-11}$. Under an independently sampled orthogonal gauge for every evaluation context, the proposed ambient-error upper bound is $4.21 \times 10^{-15}$, whereas the fixed-axis arm has relative gauge defect 1.26298. All eight registered $K = 500$ lifts pass; the largest coordinate, protected-complement and endpoint errors are $7.22 \times 10^{-16}$, $6.59 \times 10^{-17}$ and $4.51 \times 10^{-17}$, with no $K \times K$ action. This constructively qualifies the request representation and its identifiability. Because the coefficient law is linear and the paired known-truth action is noiseless, it does not establish a real factual-label estimator or a proper-score gain.

### P.4 UNPAIRED STOCHASTIC PROPER-RISK LEARNING

The next frozen stochastic construction tests learning without paired actions or coefficient labels. The estimator sees only past-context invariants, an independent $M = 64$ q0 ensemble, and one realized $H = 16$ future per origin. Scalar and directional controllers receive the same optimization tape and minimize sampled fair Energy directly. Each replicate uses 512 training, 256 validation, and 512 evaluation origins. Consecutive groups of eight origins are the inferential units; members and path coordinates are not treated as replicates. Validation activates the directional checkpoint only when its relative gain over the scalar checkpoint exceeds 0.05% and a one-sided block-$t$ lower bound is positive. No refit follows activation.

Across 32 preregistered replicates at each directional dose, activation is 0/32 under the null, 0/32 at the weak dose, 29/32 at the moderate dose, and 32/32 at the strong dose. The null false-activation Wilson upper bound is 0.1072; the strong activation lower bound is 0.8928. At the strong dose, deployed fair Energy improves q0 by 0.5128% and the scalar controller by 0.1739%. None of the 32 activated checkpoints harms the scalar comparator on its independent evaluation tape; the Wilson upper bound is 0.1072. The mean directional-coefficient RMSE is 0.1090. Gauge replay is below $6.60 \times 10^{-15}$, and all eight $K = 500$ thin lifts pass without a dense action. This closes stochastic

learnability and selective-safety questions on the constructed law. It does not establish a real-family factual gain or upgrade either scoped SOTA result.

### P.5 Noisy-history shape estimation and fail-closed deployment

The previous construction supplied the covariant shape exactly. A harder frozen $K = 500$ study requires estimating $S(c)$ from a noisy 96-step thin history. This is a distinct failure mode: matrix perturbation can rotate an estimated eigenspace sharply when the signal gap is small (Yu et al., 2015), so validation-clean directional gains need not survive deployment noise. Each of 32 replicates uses independent 512/256/512 train/validation/evaluation origins. Training and validation have Gaussian noise scale 0.20. Evaluation crosses matched Gaussian 0.20, benign Gaussian 0.35, severe Gaussian 0.80, and variance-normalized Student-$t_3$ 0.80.

Two tempting origin-level reliability proxies fail in development. Odd/even split-half shape cosine is too noisy with only 48 rows per half, while raw anisotropy rises under heavy-tailed outliers even as true-shape alignment falls. The retained estimator clips centered row radii at the within-block 0.80 quantile before forming a trace-free covariance; the operation is orthogonally invariant. A batch-level preflight uses only past contexts: it compares deployment invariant medians to training medians in training-MAD units and monitors the drop in empirical–robust shape agreement. It never accepts an evaluation future. Failure returns the scalar paths bit-exactly.

The robust checkpoint activates in 32/32 replicates (Wilson lower 0.8928). The preflight passes matched and benign noise in 32/32 cases each, while rejecting Gaussian-severe and Student-$t_3$-severe noise in 32/32 cases each. Matched deployment improves q0/scalar Energy by 0.4478%/0.1168%; benign deployment improves them by 0.3606%/0.0507%. Both severe deployments equal the scalar comparator exactly. This safety boundary is nonvacuous: under the heavy shift, the activated unguarded empirical-shape arm worsens scalar Energy by 0.5933%, whereas radial winsorization raises mean true-shape alignment by 0.0919. Gauge replay is $7.88 \times 10^{-15}$ and all eight $K = 500$ lifts pass. The result qualifies synthetic noisy-history robustness inside the registered envelope, not real-frame estimation, physical-tail realism, or distribution-free safety.

We next froze the checkpoint, estimator, activation bit and both preflight thresholds, then moved to ten contamination scenarios absent from the original development. Across 32 new replicates, low colored-AR(1), sparse-burst, heteroskedastic and coherent trace-preserving shifts pass in 32/32 cases and improve scalar Energy by 0.1031–0.1414%. Their three registered high scale/outlier counterparts reject in 32/32 cases and return scalar paths exactly. Coherent-high remains usable despite mean true-shape alignment of 0.4122, improving scalar by 0.0599%; alignment alone is therefore not a risk surrogate. The negative control reveals the boundary. A trace-preserving covariant shape reflection rotates empirical and robust estimates together. Both scales pass the unchanged preflight in 32/32 cases, yet worsen scalar by 0.0973% and 0.0958%. Thus the preflight transfers across the registered ordinary extensions but is an outlier/scale-shift screen, not a general shape-shift certificate. Gauge replay is $9.98 \times 10^{-15}$ and all eight $K = 500$ lifts pass; the failure is statistical observability, not control geometry.

**Proposition 3** (Past-only invariant observability limit)**.** *Let $T(CQ) = T(C)$ for every thin history $C$ and $Q \in O(r)$. If $R(C) \in O(r)$ is gauge covariant, $R(CQ) = Q^\top R(C)Q$, then $T(CR(C)) = T(C)$. Hence no gate measurable only through this invariant information can uniformly distinguish $C$ from $CR(C)$, although a shape-conditioned action applied to an unchanged q0 law can have different proper risk. Appendix I gives the proof and numerical boundary.*

A preregistered $K = 500$ stress test instantiates the proposition with a past-derived spectral reversal that uses no true shape, future or score. Clean and reversal both pass the unchanged preflight in 32/32 replicates, while all registered past diagnostics agree within $7.82 \times 10^{-14}$. Clean improves scalar Energy by 0.1517%; reversal worsens it by 0.0965% with 32/32 harms. The formal terminal is nevertheless *unresolved*: recomputing invariant features after the orthogonal action does not retain bit-exact scalar paths, although scalar score means agree within $5.56 \times 10^{-17}$. We retain that frozen failure rather than relaxing exact equality after seeing it.

The proposition also suggests a constructive remedy: add a relation outside the single-view invariant sigma-field. For two unit-Frobenius trace-free shape estimates, $a(S_1, S_2) = \langle S_1, S_2 \rangle_F$ is unchanged by a common gauge, but can change when a right action distorts only one view. We froze a 5-MAD cross-view gate using two independently noised past views. Clean and a moderately noisier auxiliary

view pass in 32/32 replicates and improve scalar Energy by 0.1208%. A primary-only reversal still passes the old preflight in 32/32 cases, but the relation rejects it in 32/32 and returns scalar paths exactly, avoiding 0.1060% unguarded harm. The required common-cause control applies the same reversal to both views: their relation changes by at most $1.22 \times 10^{-15}$, all 32 cases pass and all 32 are harmful. Thus an independent relation is sufficient in the registered model, not a universal certificate; real auxiliary-view availability and failure independence remain untested.

An existing real asset lets us test whether such a relation is observable without scoring a future. Ten frozen nearby-station pairs from GHCNh (NOAA National Centers for Environmental Information, 2026) define a $K = 10$ A-station/B-station pressure panel. A create-once 2022–2023 train-year study uses 56 fit, 28 split-conformal calibration, and 84 evaluation past windows. Primary-only reversal, geographic coordinate mismatch, and time mismatch are each rejected in 84/84 windows; their AUROCs against natural pairing are 1.000, 1.000, and 0.9999. Common-action relation error is $8.88 \times 10^{-16}$. The frozen terminal remains *unresolved*, however: natural-pair acceptance is 58/84 and its Wilson lower bound 0.585, below the 0.70 gate. A year-end calibration slice did not cover full-year natural regimes. Thus the real relation is strong, but a binary single-regime threshold over-abstains; conditional sensor-error independence, forecast-quality gain, and $K = 500$ transfer remain unproved.

We froze the repair before opening 2024. A full-year 2022–2023 relation reference and leave-one-year-out 95% uncertainty temperature define $w(z) = \exp[-(z/c)^2/2]$ and the action $\exp(0.05I + 0.35wS_{\mathrm{primary}})$. A circular-calendar alternative was discarded on 2022–2023 because it did not improve transport. In the one-shot 2024 mechanism evaluation, clean mean/median weight is 0.9556/0.9874 and 82/83 windows retain $w \geq 0.5$ (Wilson lower 0.9349). Primary reversal, coordinate mismatch, and time mismatch each produce 83/83 weights at most 0.05 (Wilson lower 0.9558). Common-action weights still match clean within $4.15 \times 10^{-14}$, and gauge-equivariant action error is $7.12 \times 10^{-15}$. All 19 frozen gates pass. This establishes real relational uncertainty contraction at $K = 10$, not predictive-law improvement: no future target or proper score is used, and this study does not access 2025. This local access boundary is not a global freshness certificate; see Appendix P.6.

The next study tests whether the relation improves a proper score, without opening 2024 or 2025. On 150 previously consumed 2022–2023 $K = 20$ D3U anchors, 24-fold leave-one-calendar-month-out selection compares the same base ensemble under no request, scalar spread, and a relational action applied only to A-station member anomalies. Relational Energy ratios are 0.995821 versus no request and 0.999629 versus scalar, with 91/150 and 78/150 origin wins and 18/24 month wins against both. This is a reproducible directional signal, but not a qualified controller. Sparse variogram is 1.03351 times the better baseline, exceeding the frozen 1.02 no-harm envelope; all folds also choose the upper grid boundary. The terminal is therefore STOP at 12/13 gates. We do not expand the grid, relax the gate, or proceed to 2024. The result localizes the remaining problem: one-sided marginal anomaly control can improve global Energy while damaging cross-station dependence.

This failure yields a sharper construction. Let $D$ be a predeclared graph incidence matrix and choose a thin frame $U$ with $DU = 0$. The coupled action

$$T_G = I + U(G - I)U^\top \tag{31}$$

is evaluated as $x + ((xU)(G - I))U^\top$, without a dense $K \times K$ matrix. Then $DT_G = D$ pathwise, while the available rank obeys $r \leq K - \mathrm{rank}(D)$. Ten interleaved pairs leave 10 dimensions at $K = 20$ and 250 at $K = 500$, enough for rank four. In contrast, the registered offset graph $(1, 2, 4, 8)$ is connected and leaves only the global common mode; protecting every such increment and retaining rank-four spatial control are incompatible in this linear family.

A preregistered synthetic strong-dose atlas tests this repair over 32 independent replicates. Graph-constrained Energy ratios are 0.997172/0.998877/0.999205 versus q0/scalar/one-sided, with 32/31/31 wins. Paired increments are preserved within $8.88 \times 10^{-16}$ and eight $K = 500$ thin lifts pass. The one-sided control harms pair variogram in 32/32 replicates, whereas the constrained action also improves the full sparse variogram in 32/32 against every comparator. All 16 gates pass. This is a constructive geometry certificate under a synthetic strong dose, not a learned real request or a field-effect estimate.

We then isolate real request compilation from direct outcome evaluation. Using stored 2022–2023 fit-history slices, a C16-compatible fit-snapshot reconstruction, and 150 existing GHCNh $K = 20$ D3U base ensembles ($M = 8, H = 16$), we compile the same rank-four paired-nullspace action without directly reading a target array or computing a proper score. This does not certify upstream inputs as past-only (Appendix P.6). All 150 origins respond nontrivially and 149 differ from the scalar endpoint; the relative RMS law displacement is 2.945%. Ensemble-mean, paired-contrast, thin-span, and gauge errors remain below $8.89 \times 10^{-16}$. Real-history-derived gains also pass eight $K = 500$ thin lifts using a $500 \times 4$ frame and no dense action. All 13 gates pass. This closes real-generator-base execution, not factual benefit: the experiment contains no outcome comparison and supplies no field-effect or SOTA evidence.

A fixed four-arm retrospective diagnostic then opens only the already consumed 2022–2023 targets. The graph-action Energy ratios are 0.998315/0.999253/0.999136 versus q0/scalar/matched one-sided, with 19/16/19 wins across 24 calendar months. Its full sparse-variogram ratios are 0.999836 versus q0 and 0.998895 versus one-sided; all seven existing secondary no-harm checks pass. The registered-pair score matches q0 at every origin within $5.56 \times 10^{-17}$, whereas the matched one-sided action has maximum pair-contrast error 0.2626 and worsens pair score by 0.325%. The q0 score ledger exactly reproduces C17. These results isolate coupled geometry while retaining a small factual signal, but they are post-selection development evidence because the action was designed after C17 and the targets were not fresh.

The next comparison tests whether any graph-safe rank-four action with the same dose would have produced the same improvement. Holding the base ensemble and paired-common frame fixed, we exhaust all 192 unique signed-permutation conjugations of each gain, modulo the central $\{I, -I\}$ kernel. Every orientation preserves the per-origin eigenvalues, determinant, $\|G - I\|_F$, member mean, and registered pair contrasts. The stored-history-derived identity request ranks 7/192 by mean Energy, lies 0.0358% below the 191-control median, and beats the control median in 16/24 months. Six misaligned orientations score lower, with the best ahead of identity by only 0.0050%. Thus the request is unusually favorable, not uniquely optimal. The orbit rank is descriptive rather than a confirmatory $p$-value because the target tape was already consumed when this falsification was designed.

### P.6 Covariance-matched controls and origin-conditioning ablations

**Evidence domain and protected objects.** These follow-ups reuse the same 150 consumed 2022–2023 GHCNh origins, eight-member D3U ensembles and rank-four paired-common frame from C19–C21. They introduce no new training or target year. Their member-anomaly action preserves each time step's ensemble mean, adjacent-channel pair differences and frame complement. It does not preserve the cumulative temporal endpoint of Equation 3, and is not a test of the later event/q99-protected MERRA-2 correction. All methods share the same base, frame and declared protection. We retain the historical off-diagonal Energy functional and eight companion metrics, without a new unbiased-risk claim for coupled ensembles. Source auditing subsequently identifies cross-origin interpolation in 3/150 histories (Appendix P.6); these scores do not validate strict past-only forecasting. This is not projected diffusion sampling inside a denoising process.

**Execution is reproducible by a strong generic control.** For an origin, stack the member-centered thin coordinates into $Z \in \mathbb{R}^{n \times 4}$, $n = 8 \cdot 16$, with covariance $S = Z^\top Z/n$. The original symmetric gain requests $T = GSG^\top$. We compare $ZG^\top$ with a whiten–color map $Z_{\mathrm{WC}} = ZS^{-1/2}T^{1/2}$ and an independently implemented minimum-displacement SVD construction. If $ZT^{1/2} = L\Sigma R^\top$ is a thin SVD, the latter uses $Z_{\mathrm{SVD}} = \sqrt{n}\, LR^\top T^{1/2}$, then lifts and centers the change in the same protected frame. Maximizing the corresponding trace in the covariance-constrained least-squares problem gives this solution. All 150 positive-definite cases satisfy the covariance and protection tolerances. WC changes 149/150 outputs; SVD reproduces the original graph outputs to numerical precision (maximum absolute sample difference $6.994 \times 10^{-15}$). Thus this experiment supplies no exclusive covariance-execution primitive.

At the original gain, mean Energy is 0.213666706 for q0, 0.213306647 for graph, 0.213305447 for WC, and 0.213297929 for a fixed fit-frame orientation retaining every origin's gain spectrum. SVD matches graph. Fixed orientation and WC are already slightly lower-scoring than graph; tiny dif-

ferences are not asserted to be significant. The earlier 192-orientation ranking is only a comparison within that particular orbit, not evidence that origin-conditioned direction is necessary.

**Adaptive search and a stronger static control.** After observing those results, we performed two bounded search stages, followed by a separately specified conditioning ablation. With original gain $G_i^{\mathrm{old}}$, define

$$D_i = \frac{\log G_i^{\mathrm{old}} - 0.05I}{0.35}, \qquad G_i(\beta_0, \beta_1) = \exp(\beta_0 I + \beta_1 D_i). \tag{32}$$

Graph, WC and fixed-orientation policies each receive 25 settings in stage one and 25 in stage two. Their combined grid is $\beta_0 \in \{-0.2, 0, 0.05, 0.2, 0.4, 0.5, 0.65, 0.8, 1, 1.2\}$ and $\beta_1 \in \{-1, -0.35, 0, 0.35, 1\}$. The extension was motivated by the first-stage strength-boundary winners; it was not fixed at study inception. Scalar receives 50 strengths: $-0.2 + 0.025j$ and $0.425 + (1.2 - 0.425)j/24$, $j = 0, \ldots, 24$, with $\beta_1 = 0$. These are matched configuration counts, not identical search geometries or compute budgets. Any gain eigenvalue outside $[0.25, 4]$ rejects the entire configuration without clipping or replacement. This is the exploratory C19-family range, not a change to the original KRPC request envelope.

For the new ablations, use only the 74 timestamp-labelled 2022 generators to form $A = \mathrm{mean}_{2022} D_i$, $V = A/\|A\|_F$, and $r_D = (\mathrm{mean}_{2022}\|D_i\|_F^2)^{1/2}$. The static policy substitutes $D_i^* = r_D V$ at every origin; magnitude-only substitutes $D_i^* = \|D_i\|_F V$. Each receives the same 50-point matrix grid. The first removes all origin variation of the control matrix; the second retains only its scalar norm. RMS matching avoids weakening the static control merely through mean cancellation, but does not match its eigenvalues or displacement origin by origin. Both include the isotropic part of $D_i$; neither is a pure eigenvector-only ablation. The static action still depends on the current base ensemble. Its template is target-free, but upstream frames and gains use a two-year fit snapshot, so the 2023 partition is not an independent forecasting holdout.

The first three stages contain 300 family-labelled configuration opportunities, 258 admissible: 41 each for graph, WC and fixed orientation; 50 for scalar; 43 for static and 42 for magnitude-only. Zero-direction settings intentionally duplicate outputs across families. These are not 300 independent hypotheses. All admissible samples are persisted before each stage's target scoring. Every configuration and rejection is retained; no stronger threshold or replacement candidate rescues a failed configuration.

Selection minimizes 2022 mean Energy subject to the original q0-relative secondary limits: ratios at most 1.02 for marginal CRPS, temporal-feature Energy, projected DCT CRPS, sparse variogram and quantile risk; excess at most 0.01 for coverage and tail-calibration errors. Ties use configuration order, with q0 fallback if none qualifies. The chosen configuration is then reported unchanged on the 76 consumed 2023 origins and on all 150 origins. Pooled optimistic searches are recorded separately, not substituted for that choice. No confirmatory interval or significance claim is attached to these adaptively designed comparisons.

Table 39: GHCNh retrospective policy comparison after 2022 selection. Energy reductions are relative to q0, with original origins equally weighted. Neither the pooled nor the 2023 column is fresh confirmation.

| Policy | $(\beta_0, \beta_1)$ | All reduction | 2023 reduction | 2023 gates |
|---|---|---|---|---|
| Graph | $(.50, .35)$ | $0.6990\%$ | $0.3137\%$ | Pass |
| Whiten–color | $(.50, .35)$ | $0.6999\%$ | $0.3243\%$ | Pass |
| Scalar | $(.521875, 0)$ | $0.6972\%$ | $0.3647\%$ | Pass |
| Fixed orientation | $(.50, -1)$ | $0.6875\%$ | $0.1603\%$ | Fail |
| Static template | $(.50, .35)$ | $0.7598\%$ | $0.4061\%$ | Pass |
| Magnitude-only | $(.50, .35)$ | $0.7600\%$ | $0.4009\%$ | Pass |

**What the controls rule out.** Table 39 does not support an added Energy benefit from origin-conditioned matrices. Graph's mean is 0.212173103, versus 0.212043163 for static and 0.212042941 for magnitude-only. The last two differ by only $2.23 \times 10^{-7}$ and reverse order in 2023. They retain anisotropic relational structure, so this does not establish that all relation information is unnecessary or that pure scalar calibration is sufficient. The selected fixed-orientation policy

fails 2023 marginal CRPS with ratio 1.022788 to q0; we do not replace it with the earlier $(.40, -1)$ pooled winner. For graph at $\beta_1 = .35$, increasing $\beta_0$ from .50 to .65 worsens 2022 Energy and crosses CRPS/variogram limits. This is a grid-level turning point, not a global optimum proof.

Table 40: All nine pooled GHCNh diagnostics for selected policies. Lower is better. Passing tolerated secondary gates does not mean improving each metric. These means use 150 origins, not two equally weighted years.

| Metric | q0 | Graph | Static | Magnitude-only |
|---|---|---|---|---|
| Off-diagonal Energy | .213666706 | .212173103 | .212043163 | .212042941 |
| Marginal CRPS | .176278135 | .178684202 | .178429367 | .178426866 |
| Temporal-feature Energy | .139335075 | .139179336 | .139175753 | .139208041 |
| Projected DCT CRPS | .152871399 | .150807289 | .150573898 | .150576341 |
| Sparse variogram | .067188899 | .068165485 | .067892151 | .067905032 |
| Quantile risk | .072653665 | .072027426 | .071902211 | .071900856 |
| Coverage error | .126104167 | .084826389 | .086340278 | .086326389 |
| Tail-calibration error | .080833333 | .057052083 | .057510417 | .057364583 |
| Registered pair variogram | .140994085 | .140994085 | .140994085 | .140994085 |

Table 41: Remaining selected-control diagnostics on all 150 origins, using the same nine metrics and weighting as Table 40. The fixed-orientation column is retained despite its failed CRPS gate.

| Metric | Whiten–color | Scalar | Fixed orientation |
|---|---|---|---|
| Off-diagonal Energy | .212171216 | .212177031 | .212197842 |
| Marginal CRPS | .178685144 | .178560991 | .179826930 |
| Temporal-feature Energy | .139101953 | .139173384 | .139840137 |
| Projected DCT CRPS | .150769065 | .151010445 | .150976172 |
| Sparse variogram | .068176097 | .068481746 | .067748396 |
| Quantile risk | .072032803 | .072034436 | .071870099 |
| Coverage error | .085284722 | .087180556 | .078000000 |
| Tail-calibration error | .057020833 | .058385417 | .049802083 |
| Registered pair variogram | .140994085 | .140994085 | .140994085 |

Graph retains slightly lower coverage and tail-calibration errors than the new templates; all three have worse marginal CRPS and sparse variogram than q0 (Table 40). Protection is not across-metric dominance. Independent terminal checks recompute Energy and pair scores, covariance/action geometry and selections; the other seven metrics receive finite-ledger and aggregate checks, not independent formula recomputation. The new ablation alone checks 12,900 origin–arm scores including q0, with maximum Energy error $5.00 \times 10^{-16}$ and geometry error $6.17 \times 10^{-15}$. The original frozen results remain unchanged. These controls narrow quality attribution and motivate testing a frozen policy elsewhere; they do not retroactively promote the static template.

**Quadratic spectral extension and selection instability.** After those stages, a bounded extension adds

$$G_i = \exp\Big(\beta_0 I + \beta_1 D_i + \beta_2 \Big[D_i^2 - \tfrac{\mathrm{tr}(D_i^2)}{4} I\Big]\Big).$$

The trace-free term changes spectral curvature without changing its contribution to log-determinant. Graph, static, magnitude-only, fixed orientation and WC each receive the same 36 settings: $\beta_0 \in \{.4, .5, .6\}$, $\beta_1 \in \{-.35, 0, .35\}$ and $\beta_2 \in \{-.6, -.2, .2, .6\}$. The static and magnitude variants substitute their previously defined $D_i^*$; fixed orientation retains the requested spectrum and WC the graph target covariance. Scalar receives 36 unused strengths selected at evenly spaced integer indices from a 401-point $[.35, .65]$ grid after excluding old strengths within $10^{-10}$. All 216 new settings satisfy the unchanged gain bounds. The cumulative budget is 516 labelled opportunities, 474 admissible, with 86 opportunities per family; equivalent actions are not independent hypotheses. The old 258 admitted ledgers are reused exactly. Selection still uses only 2022 Energy and its original secondary gates, without replacing failures using 2023 results.

Table 42 reports all cumulative selections. Graph now selects $(.50, .35, .20)$: its 2022 gain rises from 1.067030% to 1.081323%, while consumed-2023 gain falls from .313702% to .236948%. Every

newly selected policy has lower 2023 and pooled Energy improvement than its own prior selection; fixed orientation retains its old selection. This is selection instability on consumed data, not an independent test of overfitting. Scalar's sparse-variogram ratios are 1.021070 in 2023 and 1.020362 pooled, exceeding 1.02. Fixed orientation retains its CRPS failure. Pooled selection instead gives graph .713796%, versus static .768166%, magnitude-only .772840% and WC .714336%; these optimistic development values do not replace the 2022-selected results. The extension establishes neither a unique execution primitive nor an origin-conditioning advantage.

Table 42: Cumulative 86-opportunity selections per family. Energy reduction against q0 (%); pooled means weight all 150 origins equally. The final column gives secondary-gate status in 2023 and pooled data. All years were previously consumed, with the preprocessing limitation described below.

| Policy | 2022 | 2023 | Pooled | Gates |
|---|---|---|---|---|
| Graph | 1.081323 | .236948 | .668851 | Pass/Pass |
| Whiten–color | 1.071868 | .252333 | .671530 | Pass/Pass |
| Static | 1.141075 | .258660 | .710021 | Pass/Pass |
| Magnitude-only | 1.144692 | .258573 | .711828 | Pass/Pass |
| Scalar | 1.025897 | .345393 | .693475 | Fail/Fail |
| Fixed orientation | 1.190878 | .160318 | .687456 | Fail/Fail |

Table 43: All nine pooled diagnostics for the cumulative selections. Lower is better; passing tolerances does not mean improvement in every metric. WC denotes whiten–color and Mag. magnitude-only.

| Metric | q0 | Graph | WC | Static | Mag. | Scalar | Fixed |
|---|---|---|---|---|---|---|---|
| Energy | .213666706 | .212237594 | .212231870 | .212149628 | .212145767 | .212184981 | .212197842 |
| Marginal CRPS | .176278135 | .178903775 | .178914817 | .179458295 | .179456158 | .178725810 | .179826930 |
| Temporal Energy | .139335075 | .139228361 | .139160415 | .139464587 | .139486717 | .139214051 | .139840137 |
| DCT CRPS | .152871399 | .150877908 | .150830832 | .150826696 | .150823445 | .151038876 | .150976172 |
| Sparse variogram | .067188899 | .068136814 | .068153170 | .068271000 | .068282439 | .068557020 | .067748396 |
| Quantile risk | .072653665 | .072041711 | .072049928 | .071941825 | .071939874 | .072033483 | .071870099 |
| Coverage error | .126104167 | .083750000 | .083861111 | .081090278 | .080923611 | .085652778 | .078000000 |
| Tail error | .080833333 | .055895833 | .055750000 | .052531250 | .052583333 | .057427083 | .049802083 |
| Pair variogram | .140994085 | .140994085 | .140994085 | .140994085 | .140994085 | .140994085 | .140994085 |

An independent audit checks the 216 persisted actions, polynomial spectra, protection, target covariance and SVD equivalence, recomputes Energy and pair scores for 32,550 origin–arm entries including q0, and verifies cumulative selection. The other seven diagnostics receive finite-ledger and aggregation checks, not independent formula recomputation. No new base training or random tape is introduced. All historical results remain unchanged.

**Interpolation provenance of the consumed anchors.** The legacy loader interpolates closed gaps of at most six hours on each whole-year series before slicing origins. Reconstructing all 40 station-year inputs for 2022–2023 exactly reproduces the saved $17{,}520 \times 20$ float32 FIT array. An independent linear-gap formula has zero maximum discrepancy. Per-cell right-endpoint tracing and a separate gap-interval intersection calculation agree on all 150 origins, retaining zero-count rows.

Seven history cells in three origins (3/150) use a right endpoint at or after the origin, outside the strict history interval $[o-96, o)$. All three are in 2023 (3/76); none is in 2022 (0/74). Their UTC origins are May 21 05:00, October 24 03:00 and December 5 03:00, with two, two and three cells respectively. The latest endpoint is $o + 2$ hours. Separately, one target cell at the January 19 00:00, 2022 origin uses the endpoint $o + 16$, outside the target window $[o, o + 16)$. This is a target-definition issue, not an additional history-leakage case. Of the 189 imputed history and 36 imputed target origin–cells, only the counts just specified cross their respective right boundaries.

No counterfactual scores were computed in that provenance audit. The subsequent fixed-weight replay below leaves the quality effect at changed origins unidentified because their raw targets are incomplete. We retain all 150 origins and all reported results. The remaining 147 histories are not thereby certified as fresh or past-only: the backbone, standardizer and reference use the consumed two-year FIT, and observation-release latency has not been verified. The algebraic protection identities are distinct from this predictive-validity defect. These comparisons are retrospective

FIT mechanism diagnostics, not strict past-only forecasting validation. Future validation requires an origin-prefix preprocessing and missingness contract, target-window rules and checkpoint/input compatibility checks; changing inference preprocessing alone would not certify the old model.

A separate project's records also show QC/availability access to the same 20 stations in 2025 and selection of 120 pair-windows. The present branch's local non-access record cannot establish global non-exposure. At that audit, these records did not establish the complete quality-result access history. Later records document old-station 2025 scoring; the separate nearby evaluation is reported in Appendix P.9. No 2025 values were opened for the earlier provenance audit.

### P.7 Fixed-weight prefix-repair sensitivity

**Estimand and shared replay.** This posthoc comparison asks whether repairing the consumed histories changes quality at fixed weights, not whether retraining a prefix-safe model improves forecasting. We freeze the previous 2022-selected configurations after the 516-opportunity search, including q0 and the same-request SVD control, with no new training or hyperparameter search. The three conditions are legacy inputs/requests, repaired base inputs with legacy requests, and repaired base inputs with recomputed requests. Repair replaces the seven trailing missing history cells by the last observed value only when the trailing run is at most six hours; other unresolved histories would be rejected jointly. The backbone, standardizer and reference remain the old two-year FIT objects. This does not repair their training provenance or verify release latency.

The original initial noise was not persisted. We reconstruct its recorded PCG64 recipe with seed 29082921, persist the new tape once, and share it across conditions: $M = 8$, $L = 96$, $H = 16$, $K = 20$, batches of 15 and 20 solver steps. All 150 historical float32 forecasts replay exactly; this does not verify historical noise-byte identity. The 147 unchanged histories reproduce their forecasts exactly. Only origins 103, 135 and 144 (zero-based) change their base forecasts and requests. All $150 \times 3 \times 8$ ensemble bundles are committed before extracting future target windows.

**A missing-target identification limit.** The frozen common scoring rule requires all $16 \times 20$ raw target cells to be QC-valid, without target interpolation. All 150 inputs are admitted and forecast, but only 133 origins have complete targets: 67 in 2022 and 66 in 2023, covering all 24 calendar months. The other 17 retain eligibility and missingness records but contribute no scores or placeholder values to any aggregate. Crucially, all three changed origins are among those 17, with 2, 3 and 2 missing target cells respectively. Thus zero of three changed origins is scoreable. On the 133 eligible origins all nine scores agree exactly across the three conditions, which are not independent replicates. The quality effect at the changed origins remains unidentified, not zero. This result neither establishes harmless interpolation nor attributes the earlier negative results to preprocessing. We do not impute targets, replace origins or change the metric to recover a favorable comparison.

Table 44: Fixed configurations on the common raw-target panel. Fair full-path Energy is lower-is-better; columns pool eligible origins equally. The three input conditions coincide on these origins and are shown once. P/F denotes the original no-harm pass/fail, not statistical significance or promotion.

| Arm | All (133) | 2022 (67) | 2023 (66) | No-harm: all/22/23 |
|---|---|---|---|---|
| q0 | 0.2159791091 | 0.2244101321 | 0.2074203434 | P/P/P |
| Scalar | 0.2144282262 | 0.2219396760 | 0.2068029666 | F/P/F |
| Graph | 0.2144880896 | 0.2217606420 | 0.2071053469 | P/P/P |
| Whiten–color | 0.2144759269 | 0.2217825897 | 0.2070585572 | P/P/P |
| SVD | 0.2144880896 | 0.2217606420 | 0.2071053469 | P/P/P |
| Fixed orientation | 0.2143757881 | 0.2214619251 | 0.2071822855 | F/P/F |
| Static template | 0.2143596556 | 0.2215931194 | 0.2070165940 | P/P/F |
| Magnitude only | 0.2143542548 | 0.2215832810 | 0.2070156978 | P/P/F |

**Matched controls still limit the mechanism claim.** Graph reduces descriptive Energy against q0 by 0.6904% overall and 0.1519% in consumed 2023. Covariance-matched whiten–color has lower Energy in both partitions and also passes no-harm; graph is slightly better in 2022 only. SVD remains numerically equivalent to graph. Static and magnitude-only controls have lower pooled En-

ergy but fail the 2023 marginal-CRPS no-harm gate. Scalar fails sparse variogram in the pooled and 2023 panels; fixed orientation fails marginal CRPS in both. No-harm retains the original five relative 2% and two absolute 0.01 limits. These are descriptive comparisons on a reused panel and single reconstructed tape, with no new significance test, independent confirmation or SOTA evidence. The 133 raw-target scores must not be directly subtracted from earlier 150-origin interpolated-target scores: both the target definition and denominator differ.

Table 45: All nine pooled diagnostics on the 133 complete raw targets. Lower is better. Graph and SVD agree to the displayed precision; WC is whiten–color and Mag. magnitude-only. These are not the earlier 150-origin interpolated-target scores.

| Metric | q0 | Scalar | Graph/SVD | WC | Fixed | Static | Mag. |
|---|---|---|---|---|---|---|---|
| Energy | .215979109 | .214428226 | .214488090 | .214475927 | .214375788 | .214359656 | .214354255 |
| Marginal CRPS | .177775286 | .180291927 | .180559633 | .180573318 | .181431987 | .181082916 | .181079940 |
| Temporal Energy | .140411528 | .139859105 | .140097845 | .140022849 | .140671750 | .140186228 | .140209114 |
| DCT CRPS | .154407450 | .152391909 | .152291562 | .152237133 | .152297982 | .152144590 | .152141121 |
| Sparse variogram | .067863467 | .069225166 | .068817177 | .068834419 | .068387424 | .068937899 | .068951548 |
| Quantile risk | .073361847 | .072754884 | .072807487 | .072818369 | .072576590 | .072667287 | .072664882 |
| Coverage error | .126535088 | .085471491 | .083630952 | .083834586 | .077858709 | .080748747 | .080560777 |
| Tail error | .081355733 | .058270677 | .056813910 | .056661184 | .050364192 | .053207237 | .053265977 |
| Pair variogram | .141481096 | .141481096 | .141481096 | .141481096 | .141481096 | .141481096 | .141481096 |

**Independent terminal checks.** The audit verifies the continuous ledger, source bindings, fixed actions, finite and eight-distinct members, zero reference copies, and the mean, pair-difference, protected-complement and covariance certificates. The maximum geometry discrepancy is $4.1245 \times 10^{-15}$. Raw target extraction, all nine metric formulas, yearly/monthly denominators and paired aggregates are independently recomputed, without another model forward or random draw. Registered pair variogram is 0.1414810961 pooled for every arm; its preservation is an expected geometric invariant, not a quality improvement. This terminal audit closes the execution/reporting check, not the missing-target estimand.

### P.8 Bounded history-conditioned request learning

**Request learning on previously used outcomes.** After the disclosed 516-opportunity search and fixed-weight replay, we test whether a small learned request policy adds value beyond constant mixture weights, a static matrix, and a history-conditioned scalar. This is a bounded post-hoc follow-up, not an independent confirmation. We reuse the repaired-input condition of Appendix P.7: 150 origins, of which 67 in 2022 fit the heads and 66 in 2023 provide descriptive evaluation. The 17 incomplete-target origins remain unscorable, including all three repaired origins 103, 135 and 144. Learning a request cannot identify their missing quality effect. Only head fitting is year-separated: the frozen base model, normalizer and reference already exposed both years, and the panel was consumed by earlier searches. Observation-time replay is an explicit assumption, not an authenticated publication-time certificate.

**Four heads and matched execution mechanisms.** We keep $L = 96$, $H = 16$, $K = 20$, controllable rank $r = 4$, and $M = 8$ stored base paths. Four invariant history summaries ($\log(1 + \text{energy})$, $\log(1 + \text{increment energy})$, lag-one dependence and energy change in the projected history) are standardized using only the 67 head-fit origins. The 15-parameter conditional gate uses these features to softmax-mix three log-SPD experts: identity, a shared static expert, and an origin's history expert. For trace-free history covariance direction $D$ normalized to unit Frobenius norm (or zero when its unnormalized norm is at most $10^{-12}$), the latter is $0.50I + 0.35D + 0.20(D^2 - \mathrm{tr}(D^2)I/r)$; the static expert is its training-origin mean. Exponentiating the mixed log request gives $G$. The three-parameter constant gate learns only mixture biases. Its weights are constant, but its history expert still varies by origin; it is not a history-free action. The ten-parameter static log-SPD head instead learns one symmetric $A$ with $\log G = \log(4)\tanh(A)$, where the matrix-function hyperbolic tangent is implemented using matrix exponentials and a linear solve. The five-parameter conditional scalar uses $\log G = \log(4)\tanh(w^\top z + b)I$. All gains stay in $[1/4, 4]$. Coefficients start at zero: the two gates share uniform mixture outputs, whereas static/scalar begin at identity. Thus equal update budgets do not imply equal capacity or equal initial predictions across all four families.

Each terminal request is executed by direct graph-safe action, same-request whiten–color (WC), and minimum-displacement SVD. These share the requested covariance and original base paths, isolating request choice from execution mechanism. We retain the eight earlier controls unchanged and the untrained uniform gate, giving 21 arms in a fixed reporting order.

**Fixed optimization and reporting budget.** Each head receives exactly 1,000 AdamW updates with batch size 16, learning rate 0.01, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and zero weight decay. CPU float64, one thread and deterministic algorithms are used, with no scheduler, clipping, early stopping, best-checkpoint selection or additional hyperparameter search. The common persisted batch tape samples training origins with replacement using PCG64 and `SeedSequence([2026092101,0,update])`. The objective is the origin-mean original off-diagonal full-path Energy, normalized by $\sqrt{HK}$ and using the $M(M-1)$ repulsion denominator. No base model is retrained and no new base paths are drawn. This finite-ensemble objective does not acquire an iid unbiased-risk interpretation merely by differentiating it. No-harm is evaluated outside training: marginal CRPS, temporal Energy, DCT CRPS, sparse variogram and quantile risk must each be at most 1.02 times q0; coverage and tail errors must each be at most q0 plus 0.01. Point MAE is not used for selection. All terminal heads and failures are reported.

Table 46: All 21 arms in the bounded history-policy follow-up. Entries are Energy reductions (%) relative to q0; positive is better. q0 Energy is 0.2244101321 (2022), 0.2074203434 (2023), and 0.2159791091 (pooled). Pooled includes the fitting year and is not a generalization estimate. P/F lists no-harm pass/fail in 2022/2023/pooled order, not a significance test. D denotes direct execution. No arm is omitted or selected from this table.

| Arm | 2022 ($n = 67$) | 2023 ($n = 66$) | Pooled ($n = 133$) | Gates |
|---|---|---|---|---|
| q0 | 0.0000 | 0.0000 | 0.0000 | P/P/P |
| Old scalar | 1.1009 | 0.2976 | 0.7181 | P/F/F |
| Old graph | 1.1806 | 0.1519 | 0.6904 | P/P/P |
| Old WC | 1.1709 | 0.1744 | 0.6960 | P/P/P |
| Old SVD | 1.1806 | 0.1519 | 0.6904 | P/P/P |
| Old fixed orientation | 1.3138 | 0.1148 | 0.7424 | P/F/F |
| Old static template | 1.2553 | 0.1947 | 0.7498 | P/F/P |
| Old magnitude only | 1.2597 | 0.1951 | 0.7523 | P/F/P |
| Uniform gate | 0.8504 | 0.3656 | 0.6193 | P/P/P |
| Conditional gate D | 1.3185 | 0.1841 | 0.7779 | P/P/P |
| Conditional gate WC | 1.3145 | 0.1854 | 0.7764 | P/P/P |
| Conditional gate SVD | 1.3185 | 0.1841 | 0.7779 | P/P/P |
| Constant gate D | 1.1727 | 0.1996 | 0.7089 | P/P/P |
| Constant gate WC | 1.1650 | 0.2101 | 0.7099 | P/P/P |
| Constant gate SVD | 1.1727 | 0.1996 | 0.7089 | P/P/P |
| Static log-SPD D | 1.4338 | -0.0836 | 0.7106 | P/F/P |
| Static log-SPD WC | 1.4566 | -0.0445 | 0.7412 | P/F/P |
| Static log-SPD SVD | 1.4338 | -0.0836 | 0.7106 | P/F/P |
| Conditional scalar D | 1.4732 | -0.8402 | 0.3707 | F/F/F |
| Conditional scalar WC | 1.4732 | -0.8402 | 0.3707 | F/F/F |
| Conditional scalar SVD | 1.4732 | -0.8402 | 0.3707 | F/F/F |

**Learning does not establish the proposed cross-year advantage.** Table 46 separates a better fitting result from the proposed benefit of conditional mixture weights. Conditional direct improves pooled Energy by 0.7779% versus q0, compared with 0.6904% for the old graph. But its 2023 improvement of 0.1841% trails constant direct (0.1996%) and untrained uniform (0.3656%), all passing no-harm. The uniform arm was included to expose initialization effects, not chosen as a winner after inspecting 2023. For the learned conditional request, WC has slightly lower 2023 Energy than direct (0.2070357402 versus 0.2070385718), while direct has slightly lower pooled Energy (0.2142990794 versus 0.2143022619). Neither the conditional weights nor the graph execution mechanism has demonstrated unique cross-year value. These outcomes do not identify whether overfitting, distribution change or optimization limits explain the reversal, and do not show that history conditioning is universally ineffective.

Table 47: All nine descriptive 2023 scores for the new mixture policies, with q0 for reference; lower is better. C is conditional gate, B is constant-bias gate, U is untrained uniform. Direct and SVD agree at displayed precision. Same-request WC differences and secondary costs are retained.

| Metric | q0 | U | C D/SVD | C WC | B D/SVD | B WC |
|---|---|---|---|---|---|---|
| Energy | .207420343 | .206662084 | .207038572 | .207035740 | .207006432 | .206984634 |
| Marginal CRPS | .171237097 | .172473063 | .173843734 | .173829379 | .174269453 | .174243429 |
| Temporal Energy | .139013653 | .139420621 | .139689248 | .139675275 | .140368271 | .140332324 |
| DCT CRPS | .148177014 | .146706350 | .146985041 | .146916588 | .147004331 | .146888755 |
| Sparse variogram | .067339516 | .067803703 | .068129505 | .068139300 | .068201766 | .068218160 |
| Quantile risk | .070438789 | .070109590 | .070182995 | .070180978 | .070205491 | .070202452 |
| Coverage error | .115356692 | .090230429 | .081328914 | .082133838 | .077367424 | .077667298 |
| Tail error | .076633523 | .062665720 | .056178977 | .056202652 | .053598485 | .053196023 |
| Pair variogram | .145434282 | .145434282 | .145434282 | .145434282 | .145434282 | .145434282 |

**Negative controls and secondary costs.** Tables 47 and 48 show the nine-score 2023 profiles. The learned static log-SPD direct request worsens Energy by 0.0836% and fails marginal-CRPS no-harm; its WC version also fails marginal CRPS. Conditional scalar worsens Energy by 0.8402% and fails marginal CRPS and sparse variogram under all three mechanisms. It already fails sparse variogram on the fitting year despite lowering Energy. Old scalar fails 2023 sparse variogram; old fixed-orientation, static-template and magnitude-only fail marginal CRPS. These are terminal scientific negative outcomes, not reasons to retry or discard arms. Pair-variogram preservation is an expected invariant, not a quality gain. No new uncertainty or significance estimate was prespecified or added, and eight coupled paths are not eight independent studies.

Table 48: Descriptive 2023 scores for the learned non-mixture controls. S is the genuinely static log-SPD request; R is conditional scalar. SVD matches direct at displayed precision; all three scalar mechanisms agree.

| Metric | q0 | S D/SVD | S WC | R D/WC/SVD |
|---|---|---|---|---|
| Energy | .207420343 | .207593800 | .207512611 | .209163173 |
| Marginal CRPS | .171237097 | .175301293 | .175326140 | .176744173 |
| Temporal Energy | .139013653 | .141055888 | .140972176 | .140464259 |
| DCT CRPS | .148177014 | .147670252 | .147626456 | .148602264 |
| Sparse variogram | .067339516 | .068044049 | .068072900 | .069824225 |
| Quantile risk | .070438789 | .070117618 | .070090194 | .070972283 |
| Coverage error | .115356692 | .073169192 | .072553662 | .089157197 |
| Tail error | .076633523 | .047537879 | .047230114 | .052154356 |
| Pair variogram | .145434282 | .145434282 | .145434282 | .145434282 |

**Execution audit and stopping boundary.** One admitted CPU execution completed all 4,000 updates in 23.895 seconds. An earlier resource-admission rejection occurred before any scientific directory or updates. The completed experiment has 517 consecutive committed artifacts and saves optimizer states every 20 updates. An independent audit checks source identity, optimizer counters and moments, NumPy reconstruction of requests/actions, all nine scores, origin-equal annual and pooled means, and the no-harm flags. It does not rerun the optimization fit. All 21 arms have eight distinct finite members and zero exact copies among 16,050 training-reference paths. Maximum geometry error is $5.28188 \times 10^{-15}$ and direct/SVD discrepancy is $2.88264 \times 10^{-15}$. Complete three-partition scores and flags are retained in the terminal report. Audit completion does not repair source exposure or missing outcomes, and does not establish SOTA. We stop this bounded experiment without retuning on 2023 or opening new years.

### P.9 Frozen nearby-station evaluation with incomplete annual coverage

**A fixed deployment on nearby stations.** This study uses 20 nearby stations in ten fixed protected pairs, distinct from the consumed-outcome comparison in Appendix P.8. The base was fitted on old-station 2022–2023 data. New-source normalization, reference and policy statistics use 220 eligible 2022–2023 FIT origins. Four request heads each receive 1,000 updates with batch size 16 before deployment is frozen. Evaluation uses these terminal weights and statistics without refitting or checkpoint selection; it adds zero training updates. Constant mixture weights still act on a history-varying expert, whereas the static log-SPD arm supplies a history-free request. Conditional direct is the fixed candidate, not the best arm chosen after evaluation. Uniform and constant are the primary scientific controls; WC and SVD share the conditional request and test its execution mechanism.

The calendar starts January 5, 2025 at 00:00 UTC, advances every 25 hours, and contains 346 scheduled origins with $L = 96$, $H = 16$, $K = 20$, $r = 4$ and $M = 8$. Station order, clock, QC and eligibility rules are fixed before observation acquisition. Only internal history gaps of at most six hours may be repaired, with the right donor strictly before the origin; targets require complete raw observations. No target interpolation, station substitution or clock search is performed. Acquisition covers the fixed 20 files and preserves all source-only exclusions.

**Primary estimand unavailable.** Only 55 of 346 origins qualify; 291 fail history and/or raw-target availability. January–December eligible counts are $1, 0, 1, 4, 7, 0, 13, 14, 3, 7, 2, 3$. February and June are empty, so the registered twelve-month-equal annual primary estimand is not estimable. We do not replace it by a ten-month mean, produce its annual HAC intervals, or delete months to obtain a favorable comparison. The prespecified secondary estimator averages three tapes within each origin and then weights the 55 eligible origins equally. It is conditional on observability, not an unbiased estimate for the missing weather. Three tapes, eight members, twenty channels and two unit systems do not constitute independent repetitions of the weather panel.

**Shared-path comparison and secondary results.** Each replicate uses a persisted base ensemble with addressed seed `[2026092201,2,2025,ordinal,tape]`, with tapes indexed 0–2. All eight arms reuse it without additional random draws, and every sample is committed before numerical target loading for scoring. This ordering prevents score-driven resampling; acquisition has already exposed observations and is not itself target blindness. There are 165 base ensembles and 1,320 arm-score bundles, not 1,320 independent cases. Table 49 reports all arms in FIT-standardized and hPa units. Tables 50 and 51 preserve all nine endpoint means.

Table 49: Nearby 2025 origin-equal secondary Energy; the annual primary is not estimable. Lower is better. P/F denotes descriptive tolerance pass/fail in both units, not a statistical test.

| Arm | Standardized | hPa | Tolerances |
|---|---|---|---|
| q0 | 0.191354363 | 1.519179428 | P/P |
| Uniform | 0.190698899 | 1.513265912 | P/P |
| Conditional D | 0.189890705 | 1.506729867 | P/P |
| Constant D | 0.190965737 | 1.515011538 | P/P |
| Static log-SPD | 0.191973621 | 1.522890533 | F/F |
| Scalar D | 0.189907954 | 1.506301186 | P/P |
| Conditional WC | 0.189893895 | 1.506787358 | P/P |
| Conditional SVD | 0.189890705 | 1.506729867 | P/P |

Conditional direct improves secondary Energy over q0 by 0.764894%/0.819493% (standardized/hPa), over uniform by 0.423806%/0.431917%, and over constant by 0.562945%/0.546641%. Against same-request WC the improvement is only 0.001680%/0.003815%. Against conditional scalar it is 0.009083% in standardized units but $-0.028459\%$ in hPa, so scalar is slightly better in physical units. SVD reproduces the direct scores. These are descriptive differences, not established superiority or evidence that the matrix action is irreplaceable.

The candidate passes the frozen descriptive no-harm tolerances in both units, despite slightly worse marginal CRPS and sparse variogram. Five relative endpoints may rise by at most 2% versus q0, and two calibration errors by at most 0.01. This is not statistical noninferiority, improvement on every metric, or protection at every origin. Static log-SPD fails: its marginal CRPS worsens by

Table 50: All nine nearby-2025 secondary scores in FIT-standardized units. U: uniform; D: conditional direct; C: constant; S: static log-SPD; Sc: scalar; WC/SVD: same-conditional-request execution controls. Lower is better.

| Metric | q0 | U | D | C | S | Sc | WC | SVD |
|---|---|---|---|---|---|---|---|---|
| Energy | 0.191354 | 0.190699 | 0.189891 | 0.190966 | 0.191974 | 0.189908 | 0.189894 | 0.189891 |
| Marginal CRPS | 0.159697 | 0.160935 | 0.160636 | 0.162692 | 0.165170 | 0.162274 | 0.160628 | 0.160636 |
| Temporal Energy | 0.152440 | 0.152285 | 0.152354 | 0.153000 | 0.153693 | 0.153311 | 0.152261 | 0.152354 |
| DCT CRPS | 0.139593 | 0.138209 | 0.137961 | 0.138336 | 0.139562 | 0.138698 | 0.137990 | 0.137961 |
| Sparse variogram | 0.057219 | 0.057513 | 0.057394 | 0.057878 | 0.058702 | 0.057947 | 0.057388 | 0.057394 |
| Quantile risk | 0.065271 | 0.064998 | 0.064698 | 0.065097 | 0.065574 | 0.064717 | 0.064687 | 0.064698 |
| Coverage error | 0.107216 | 0.088239 | 0.083415 | 0.078744 | 0.076484 | 0.074905 | 0.083624 | 0.083415 |
| Tail error | 0.074773 | 0.062320 | 0.060634 | 0.053759 | 0.048362 | 0.052879 | 0.060597 | 0.060634 |
| Pair variogram | 0.103776 | 0.103776 | 0.103776 | 0.103776 | 0.103776 | 0.103776 | 0.103776 | 0.103776 |

Table 51: All nine nearby-2025 secondary scores under the hPa evaluation representation; normalized and dimensionless endpoints retain their definitions. Arm abbreviations follow Table 50. Lower is better.

| Metric | q0 | U | D | C | S | Sc | WC | SVD |
|---|---|---|---|---|---|---|---|---|
| Energy | 1.519179 | 1.513266 | 1.506730 | 1.515012 | 1.522891 | 1.506301 | 1.506787 | 1.506730 |
| Marginal CRPS | 1.235753 | 1.245160 | 1.242707 | 1.258686 | 1.278066 | 1.255269 | 1.242659 | 1.242707 |
| Temporal Energy | 0.152440 | 0.152285 | 0.152354 | 0.153000 | 0.153693 | 0.153311 | 0.152261 | 0.152354 |
| DCT CRPS | 1.103412 | 1.092850 | 1.090503 | 1.093717 | 1.103011 | 1.095467 | 1.090711 | 1.090503 |
| Sparse variogram | 0.455143 | 0.457932 | 0.456729 | 0.461136 | 0.468939 | 0.461664 | 0.456677 | 0.456729 |
| Quantile risk | 0.505433 | 0.503188 | 0.500754 | 0.503876 | 0.507699 | 0.500758 | 0.500680 | 0.500754 |
| Coverage error | 0.107216 | 0.088239 | 0.083415 | 0.078744 | 0.076484 | 0.074905 | 0.083624 | 0.083415 |
| Tail error | 0.074773 | 0.062320 | 0.060634 | 0.053759 | 0.048362 | 0.052879 | 0.060597 | 0.060634 |
| Pair variogram | 0.765180 | 0.765180 | 0.765180 | 0.765180 | 0.765180 | 0.765180 | 0.765180 | 0.765180 |

3.427203%/3.424081% and sparse variogram by 2.591356%/3.030994%. All other arms pass these descriptive tolerances. Pair variogram is preserved to numerical precision; point MAE does not enter selection.

**Audit and scope.** The independent panel audit reconstructs hourly values from saved source chunks and checks all 346 history/target decisions. The terminal audit checks 353 continuous artifacts, source bytes, addressed noise, requests, actions, geometry and raw-unit certificates, all scores and an independent aggregation. All ensembles contain eight distinct paths with zero exact matches to the finite FIT reference library. Maximum geometry error is $5.224\times10^{-15}$. Noise is reconstructed only for verification; the audit does not rerun the base forecast or the optimization fits. Finite-library zero-copy checks do not establish global non-memorization. The scientific evaluation takes 28.723 seconds with shared-base batching, excluding data acquisition, engineering and independent auditing.

This is observation-time replay, not authenticated real-time availability: publication latency is unknown. Prior project evaluations have exposed 2025, so the calendar year is not globally fresh; bounded source checks cannot establish global non-use. Earlier prefix-repair effects remain unidentified. The present result supports a limited-coverage fixed-deployment description, not an annual confirmation, a generally superior learning procedure or SOTA. We retain the missing denominator and all controls without retuning on this evaluation or substituting stations to fill empty months.

### P.10 Detailed interpretation and limitations

Joint time-series generators have distinct roles as predictive laws and as sources of controllable paths. A single scalar objective can confuse these roles. By Theorem 1, a request that changes the law should not be rewarded for predicting an unrequested future. The promotion contract therefore qualifies the identity law first. KRPC then adds a separate, certifiable action axis. This ordering is part of the method.

The evidence follows the same chain. Known-truth corruptions qualify what the endpoints can detect. Outer-unit reduction prevents repeated paths from manufacturing confidence. The selective policy refuses unsupported frames and requests. The frozen EEG confirmation shows that predictive quality, calibration, non-copying, and exact control can pass together. The USGS experiment supplies a separate zero-request positive test: without attributing the winner's architecture to us, a task-specific fail-closed protocol promotes published D3U from a high-dimensional frozen roster. OISST and PTB-XL supply the counterweight: exact geometry cannot rescue a weak base law, and better point prediction can destroy ensemble quality. Both the positive and negative results test the separation between generation quality and control.

The post-outer USGS mechanism experiment completes the intended separation in the opposite direction: a frozen high-quality q0 law can support a learned request-conditioned execution layer whose gain over a parameter-matched unrequested arm is measured without consulting a future target. It is an operator result, not a second quality comparison.

The subsequent chronological scalar-policy experiment falsifies a simple shortcut. Context-conditioned isotropic rescaling yields only a 0.0288% fit-only improvement and no advantage meeting the practical gate. This localizes the open problem to factual request construction rather than high-dimensional execution.

The subsequent known-truth atlas closes the representation part of that open problem: a scalar invariant law can modulate a covariant context shape and produce the same ambient action under every orthogonal thin-frame gauge. Its scalar-only and fixed-axis controls show that direction and equivariance are both necessary in the constructed law. The stochastic extension removes the exact-action labels and learns the two coefficients by proper-risk ERM. Its null-to-strong activation curve and exact scalar fallback show that the extra directional degree of freedom can be selected rather than always deployed. The noisy-history extension then estimates the covariant shape from past observations and exposes why naive deployment is unsafe: severe observation shift can turn a selected directional arm harmful. Radial winsorization plus a past-only batch preflight preserves the registered matched/benign gains and closes both severe scenarios to the exact scalar path. The remaining empirical question is therefore narrower: whether a released real past-only context supports the same frame qualification, preflight separation and proper-score gain under natural distribution shift.

The unseen-contamination confirmation narrows it again. The unchanged preflight transfers across colored, burst, heteroskedastic and coherent extensions, but its two observables can agree while both estimators share the same wrong orientation. A real deployment therefore needs an estimator-external orientation witness or calibrated abstention; more tuning of the same agreement statistic cannot certify this failure mode.

Proposition 3 makes the stronger limit explicit: if the new witness is measurable only through the same right-orthogonal invariant information, a covariant right action can preserve it pathwise. Uniform repair requires information outside that sigma-field or explicit uncertainty-aware contraction toward scalar.

The cross-view experiment supplies the minimal constructive counterpart. An independent auxiliary history can expose a primary-only orientation error without naming thin axes, but a common-cause distortion preserves the relation.

The GHCNh K10 study shows that a predefined real station-pair relation can be highly discriminative, while also exposing the next failure: a conformal cut calibrated on one seasonal regime rejects too many natural windows in another. The frozen cross-year repair then turns that failure into a continuous action: 2024 natural windows retain high anisotropic weight, while all registered observable mismatches contract to the scalar endpoint. The practical assumption has therefore moved rather

than disappeared. A real deployment must still justify sufficiently different failure mechanisms; a common distortion of both views remains invisible.

The fit-only K20 bridge then finds a small, consistent Energy benefit but fails its sparse-variogram no-harm gate. Thus the real relation is decision-relevant, yet the exact one-sided action is rejected: orientation evidence alone does not preserve the joint spatial law. The graph-constrained repair adds the missing contract: declare protected contrasts through $D$ and spend control rank only in $\mathrm{null}(D)$. Its rank ceiling makes the trade-off explicit rather than hiding dependence damage inside a secondary score. The target-free GHCNh execution then verifies that this is not only a synthetic tensor interface: a real-history request moves an existing D3U law while preserving both its conditional ensemble mean and the declared pair contrasts. It still cannot tell whether that movement is forecast-useful. The matched-action retrospective diagnostic provides a stronger mechanism check. Coupling alone removes the registered pair-score damage and improves full sparse variogram, while the same request retains a small Energy advantage. Because those targets were already consumed, the result motivates a fresh test rather than replacing one.

The exhaustive orientation falsification narrows the mechanism claim further. Most spectrum-matched graph-safe misalignments score worse than the fixed request, but six score slightly better. This is descriptive specificity within one orbit, not evidence of necessary origin conditioning. The stronger SVD, fixed-template and adaptive-search controls in Appendix P.6 remove that stronger interpretation: the action can be reproduced, and static relational structure retains the observed Energy benefit. All subsequent tuning is disclosed as development.

Four bounded request-head fits (Appendix P.8) improve training-year Energy without establishing added consumed-2023 benefit from conditional mixture weights. Constant-weight and even untrained mixtures outperform the conditional gate there. This is counterevidence to the proposed advantage, not a proof that all history-conditioned policies are ineffective.

The MERRA-2 study makes the distinction operational under base-model misspecification. Matching calibration opportunity prevents unused generic capacity from being mislabeled as a controller innovation. Conversely, a stronger proposal need not yield a larger control gain: D3U's lower starting Energy coexists with an unresolved Q2 Energy increment. Positive spatial and dynamic effects are evidence about those properties, not permission to ignore tail or temporal costs. This supports the attribution design, not a universal improvement theorem. The completed copula extension reinforces this boundary: TACTiS-2 gains 0.63446% Energy over its own Q1 while worsening increment Energy by 4.11267%. Its four neural architectures share one environmental family and one exposed development year; they are not independent replications.

The empirical ceiling remains limited. The controlled positive result covers one EEG condition, one fit per neural method, and a four-method panel whose Diffusion-TS comparator is a disclosed local adaptation. Subject bootstrap intervals condition on the fixed checkpoints. USGS adds a separate data family for zero-request promotion and a nine-arm roster, but the promoted generator is D3U, not our proposed candidate, and physical-tail realism remains unqualified. The complete-record station screen is not representative sampling, and the neural comparisons use one terminal fit under method-specific frozen budgets rather than compute-matched tuning; uncertainty therefore conditions on those checkpoints.

The USGS bootstrap preserves the eight overlapping origins within each calendar month but treats the 48 months as resampling blocks; residual dependence across adjacent months is not explicitly modeled, and the four yearwise rankings are descriptive robustness checks rather than additional inferential units.

The USGS learned-control experiment was frozen only after the outer result was known and required two disclosed implementation-only replays for unsupported MPS eigensystem calls before any operator metric was accessed. It therefore supports fit-only mechanism and scalability claims, not a confirmatory factual controller or an outer-quality upgrade.

The later factual-policy experiment is also post-outer and uses a D3U backbone trained across its controller evaluation years. Its chronological split isolates controller adaptation, but cannot establish independent forecasting generalization. The negative scalar result is terminal and is not permission to search a richer request on the same evaluation tape.

Likewise, the post-C16 GHCNh K20 bridge reuses 2022–2023 fit anchors for controller development. Its LOMO Energy gain is not confirmation; the frozen sparse-variogram failure closes the exact A-only specification, and neither 2024 nor 2025 is opened as a rescue tape in that A-only study.

The subsequent graph-constrained atlas is deliberately synthetic and uses a registered strong dose. Its 0.283% q0 improvement and exact pair protection validate a construction, not a natural effect size. Which graph should be protected and whether a nonzero nullspace request improves a fresh real predictive law remain open. The follow-up real-base certificate uses an existing $M = 8$ D3U ensemble without direct target-array access; it establishes execution geometry, not past-only preprocessing, real proper-risk improvement, independent-family replication, or real $K = 500$ transfer.

The subsequent C20 comparison does use outcomes, but only the 2022–2023 targets consumed before the graph action was designed. Its favorable Energy and dependence diagnostics are therefore post-selection development results. They cannot estimate fresh-family effect size or support confirmatory inference. The four-head follow-up retains these exposures. Its pooled score includes head-fitting outcomes; eight coupled paths are not eight independent studies. No new significance estimate is attached, and the observed uniform-gate result cannot become a retroactively selected confirmation winner (Appendix P.8).

The C21 orbit reuses the same consumed outcomes. Its 7/192 rank is therefore a descriptive falsification statistic, not a randomization $p$-value: orientation exchangeability is not a natural sampling assumption, and the orbit was frozen only after the graph-repair problem was visible. The six better misalignments also rule out unique-optimum language. A fresh family with a predeclared identity/scalar/graph comparison remains necessary.

The directional studies remain synthetic. The first is deliberately easy to identify: its coefficient law is linear in four recovered invariants, its paired action is noiseless, and its context covariance contains the exact covariant shape. Its near-zero error is a constructive consistency check. The second removes paired labels and uses stochastic realized futures, but retains the same correctly specified law and exact context shape. Its 0.1739% strong-dose advantage over the scalar controller is an operating-characteristic effect, not a real-data estimate; weak-dose nonactivation also shows finite power near the detection boundary. The third adds noisy 96-step shape estimation and registered Gaussian/Student-$t_3$ shifts, but the noise law, thresholds and coefficient law remain synthetic and correctly specified. Its batch gate is not a per-origin safety certificate, and four benign replicates still have directional score worse than scalar even though the pooled benign effect is positive. The fourth keeps that gate fixed across unseen synthetic contaminations and confirms both ordinary-family transfer and a deliberately adversarial shared-bias blind spot. The reflection corruption uses known true shape only to generate a falsification family, never for fitting, gating or action, and is not asserted to be a natural noise law. These studies do not establish real-frame qualification, unknown-shift safety, non-Gaussian physical tails, or fresh-family S/V/outer.

There is still no representative field-wide panel, qualified raw-law envelope, or NVIDIA comparison. The K32/K128 atlas covers ten named failure families, not every possible frame error. The known-truth request panel covers scalar, rotated SPD, and target-covariance actions through rank four, whereas the real EEG request set remains limited to three scalar gains. Deterministic KRPC certificates establish numerical and geometric correctness, not causal meaning or physical benefit. These boundaries are part of the claim, not deferred qualifications.

Within that boundary, the contribution is an ordered contract for conditional law, uncertainty, novelty, request geometry, and refusal, rather than another backbone. It applies above diffusion, flow, copula, or empirical sources. USGS extends zero-request promotion to a second data family and broader roster. Linking the same independently leading generator to qualified real-data control, and testing a representative field-wide panel, remain future evidence.

# Q COMPLETE CLAIM–EVIDENCE MAP

Table 52: Claim–evidence map. Each conclusion inherits the uncertainty of its declared evidence class.

| Authorized claim | Evidence class | Excluded scope or remaining uncertainty |
|---|---|---|
| Quality and requested control require separate targets | Strict-propriety theorem for Energy risk | Requires finite first moments and a request that changes the evaluated law |
| Stable-frame measurement policy is qualified | Known-truth $K = 32$ and $K = 128$ atlas | Ten named frame families; no real-data latent-span qualification |
| Selective KRPC is fail-closed in its registered envelope | 5,400 known-truth requests plus deterministic certificates | SPD requests, $r \leq 4$ statistical qualification, and registered margins only |
| Candidate leads the historical EEG panel | Frozen 61-subject D25 comparison | Whole-record normalization uses future increments; one condition, one fit per method, four primary methods and fixed checkpoints; not past-only forecasting |
| Published D3U leads the frozen $K = 500$ roster | USGS development, one-shot 2021 qualification, and 2022–2025 nine-arm outer confirmation | Validates task-specific zero-request promotion, not the EEG conjunction, our backbone, or our controller; raw-tail and broad cross-family claims remain closed |
| Learned operator execution does not imply a useful factual policy | K500 fit-only operator test plus chronological isotropic-controller falsification | The native fiber executes requests, but the annual-label scalar policy fails its practical gate; anisotropic fresh-family control remains untested |
| Gauge-equivariant directional requests are identifiable without named axes | Disjoint K500 constructive known-truth train/evaluation atlas | The two-coefficient representation passes; real factual supervision, quality benefit and fresh-family confirmation remain untested |
| Directional coefficients are learnable without paired exact actions | Preregistered K500 stochastic proper-risk operating characteristics | Synthetic unpaired learning and selective activation pass; real-frame estimation, distribution shift and fresh S/V/outer remain untested |
| Noisy-history directional deployment can fail closed | Preregistered K500 estimated-shape robustness operating characteristics | Matched/benign operation and two registered severe shifts pass; real-frame qualification, unknown shifts, physical tails and fresh S/V/outer remain untested |
| Unseen contamination transfers only partially | Preregistered ten-scenario K500 confirmation and adversarial negative control | Four ordinary mechanisms transfer, but a shared-bias shape reflection passes both diagnostics and harms scalar; real and distribution-free robustness remain closed |
| Past-only invariant repair has an observability limit | Exact proposition plus preregistered K500 stress test | All substantive stress outcomes support the counterexample, but formal empirical status remains unresolved on one scalar bit-exact gate; external anchors remain untested |
| An independent past relation can restore orientation observability | Preregistered K500 two-view operating characteristics plus common-mode negative control | Primary-only repair passes, but a shared distortion of both views remains invisible; real failure independence remains untested |
| A real paired-station panel carries relational information | Preregistered GHCNh K10 train-year past-only study | Three mismatch arms separate nearly perfectly, but clean retention misses its frozen gate; formal availability remains unresolved and forecast quality is untested |
| Real relational uncertainty can contract directional action | Preregistered GHCNh K10 2022–2023 fit and one-shot 2024 past-only mechanism evaluation | Clean retention and three observable mismatch contractions pass, but common-mode distortion, sensor independence, forecast quality, K500 transfer and SOTA remain unproved |
| Real relational information need not yield a safe factual controller | Post-C16 GHCNh K20 fit-only 24-month LOMO development | Energy improves over q0 and scalar, but sparse variogram violates the frozen no-harm gate; the exact A-only action stops before 2024/2025 and supports no SOTA claim |
| Graph-constrained control repairs declared relation geometry | Known-truth atlas, consumed-FIT controls and fixed-weight replay (Appendices P.6, P.7) | SVD reproduces the action. All 3 repaired origins lack complete targets; 133 eligible origins do not identify their quality effect. For the old fixed request, WC has lower pooled Energy. No strict past-only validation or fresh confirmation |
| Control correctness and point accuracy cannot proxy for law quality | Prospective OISST and consumed PTB-XL negative studies | Falsifies the stated proxies; does not rank the field or establish universal failure |
| Proposal quality, calibration and control have distinct contributions | MERRA-2 matched Q0/Q1/Q2 development on four support-native backbones | D3U has the lowest Energy mean but no established nearest-rival win; TACTiS-2 Q2 gains Energy while harming dynamics; development does not establish headroom saturation |
| Protected correction can improve held-out generation scores while harming tails | Separately frozen MERRA-2 2022 selection and locked 2023 CSDI Q2/Q1 confirmation | Energy, CRPS and increment gains coexist with q99 harm; spatial and tail Brier are unresolved; duplicate questions are one numerical pair, not whole-roster or cross-family SOTA |

Table 53: Claim–evidence map, continued: recent temporal comparisons and a failed method-entry study. The subsequent strong-backbone attribution is reported in Appendix R. Completion and geometric certificates do not override a failed quality gate.

| Authorized claim | Evidence class | Excluded scope or remaining uncertainty |
|---|---|---|
| EEG primary-comparator gains persist under prefix-only calibration | Post-hoc matched retraining on the same 61 subjects; 1,000 updates per trained configuration; Appendices F.1–F.2 | Historical off-diagonal Energy decreases 5.87–8.51% against the original three comparators and 5.73% against 100-step TimeGrad. Finite-ensemble Energy gives 4.90% against the latter; half-mix is better on that score (Appendix F.2.1). Single initialization; no fresh confirmation or SOTA |
| Event-preserving correction can transfer small Q1-relative gains | Locked MERRA-2 2024 comparison: 192 origins, three replicates, 21 configurations; Appendix N | Six variable metrics improve over Q1, but four contrasts support degradation against the strong Q1-anchored reference; Energy ranks 8/21; no whole-roster superiority |
| A target-free gate can improve a mature protected expert | Locked MERRA-2 2025 one-shot: 192 origins, three replicates, 22 configurations; Appendix O | Internal conjunction passes with 0.0600% Energy reduction; rank 3/22 and two better affine means exclude the all-roster claim; the exact-event subgroup was not a separately registered SOTA estimand |
| Bounded request learning can improve fit without establishing added cross-year value | Four heads, 1,000 updates each; 21-arm consumed GHCNh comparison, Appendix P.8 | Conditional 2023 Energy gain 0.1841% trails constant 0.1996% and untrained uniform 0.3656%. Same-request WC is slightly better there; SVD is equivalent. Static/scalar fail no-harm. Pooled gain includes fitting data; no fresh confirmation |
| Frozen request policies can yield small descriptive nearby-station gains | Frozen 2025 eight-arm evaluation; 55/346 origins, three tapes, $M = 8$; Appendix P.9 | Empty February/June preclude the annual primary estimate. Secondary Energy improves 0.7649%/0.8195% over q0, but nearly ties WC and loses to scalar in hPa; SVD equivalent. Missingness and prior exposure exclude fresh confirmation or superiority |
| Exact controls need not pass selection quality gates | Completed GPM IMERG 2021 selection with a disclosed post-unsealing execution amendment; Appendix W | Energy improves 0.013550%, but spatial variogram worsens 0.008271% and entry fails; subsequent adaptive fit work is not fresh confirmation; 2022–2024 remain unopened |

## R STRONG-BACKBONE ATTRIBUTION ON GHCNH

**Attributing gains on a stronger generator.** This comparison tests whether a protected matrix controller improves the generated path distribution beyond backbone adaptation and simple calibration. The completed comparisons are post-hoc development, not fresh confirmation. The 20-station re-panel uses 264 FIT origins from 2022–2023 and 106 eligible out of 347 scheduled 2024 origins, with monthly counts $(4, 3, 8, 18, 5, 22, 10, 4, 12, 7, 5, 8)$. All months are represented, but inference does not cover missing origins or the unconditional annual population. Histories have length 96 and forecasts length 16. Three sampling tapes of $M = 8$ members are averaged within each origin, then origins within each month, then the 12 months equally. Origin-equal averages are secondary.

**Backbone matching precedes attribution.** CSDI starts from its saved 50,200-update source checkpoint and receives a fixed 1,000-update, batch-16 adaptation on the new panel's FIT data. There is no extra hyperparameter search. Its month-equal Energy is 0.22907767/1.81730892 (standardized/hPa), versus 0.23828319/1.87584626 for the previously frozen conditional-gate system: 3.8633%/3.1206% lower. Before adaptation the same CSDI source scores 0.40506247/3.17747333; omitting adaptation would reverse this descriptive conclusion. The old gate retains lower coverage and tail calibration errors. This comparison neither matches total backbone compute nor establishes optimization saturation.

**Four fits, seven arms, one shared backbone.** All plugins reuse the adapted CSDI and its cached ensembles. A conditional scalar head has five parameters, a static log-SPD head ten, and two nested

conditional heads fifty each. Each receives the same 1,000 batch-16 updates, AdamW with learning rate 0.01 and zero weight decay, and off-diagonal Energy on the same FIT inputs. No head uses evaluation targets for gradient updates, early stopping, checkpoint selection or additional search. D denotes direct execution and W denotes whiten–color; the first letter denotes the training executor and the second the evaluation executor. DD is the pre-designated principal candidate, not the retrospectively best arm. The four fits yield q0, scalar, static, DD, DW, WD and WW. All 318 base ensembles and 2,226 arm ensembles are reused/generated without additional backbone sampling. This saves compute without cutting any comparison head's update budget.

Table 54: Complete seven-arm Energy attribution. Entries are percentage reductions against the same adapted CSDI q0; higher is better. No post-hoc winner replaces the designated DD candidate.

| | Month-equal (primary) | | Origin-equal (secondary) | |
|---|---|---|---|---|
| Arm | Standardized | hPa | Standardized | hPa |
| q0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| scalar | 3.4211 | 3.4351 | 3.0981 | 3.0774 |
| static | 3.4719 | 3.4505 | 3.2754 | 3.2259 |
| DD | 3.3811 | 3.3755 | 2.9581 | 2.8988 |
| DW | 3.4039 | 3.4000 | 2.9856 | 2.9306 |
| WD | 3.3288 | 3.3302 | 2.8900 | 2.8328 |
| WW | 3.3526 | 3.3545 | 2.9172 | 2.8640 |

Table 55: All nine month-equal metrics, FIT-standardized. Lower is better. No adverse metric or arm is omitted.

| Metric | q0 | scalar | static | DD | DW | WD | WW |
|---|---|---|---|---|---|---|---|
| Energy | 0.229078 | 0.221241 | 0.221124 | 0.221332 | 0.221280 | 0.221452 | 0.221398 |
| Marginal CRPS | 0.179372 | 0.180229 | 0.180033 | 0.180888 | 0.180915 | 0.181213 | 0.181252 |
| Temporal Energy | 0.143050 | 0.142067 | 0.141937 | 0.142451 | 0.142280 | 0.142624 | 0.142415 |
| DCT CRPS | 0.152789 | 0.150642 | 0.150782 | 0.150872 | 0.150906 | 0.150961 | 0.150996 |
| Sparse variogram | 0.049033 | 0.048826 | 0.048829 | 0.048804 | 0.048805 | 0.048867 | 0.048862 |
| Quantile risk | 0.076463 | 0.073365 | 0.073272 | 0.073533 | 0.073552 | 0.073601 | 0.073622 |
| Coverage error | 0.203922 | 0.110632 | 0.107886 | 0.109786 | 0.109589 | 0.109550 | 0.108967 |
| Tail error | 0.129076 | 0.077455 | 0.076718 | 0.078108 | 0.078229 | 0.077396 | 0.077849 |
| Pair variogram | 0.013455 | 0.013455 | 0.013455 | 0.013455 | 0.013455 | 0.013455 | 0.013455 |

Table 56: All nine month-equal metrics under the raw hPa representation. Normalized feature and calibration metrics retain their original definitions. Lower is better. No adverse metric or arm is omitted.

| Metric | q0 | scalar | static | DD | DW | WD | WW |
|---|---|---|---|---|---|---|---|
| Energy | 1.817309 | 1.754883 | 1.754603 | 1.755966 | 1.755521 | 1.756788 | 1.756347 |
| Marginal CRPS | 1.394205 | 1.400221 | 1.399156 | 1.404914 | 1.405135 | 1.407208 | 1.407550 |
| Temporal Energy | 0.143050 | 0.142067 | 0.141937 | 0.142451 | 0.142280 | 0.142624 | 0.142415 |
| DCT CRPS | 1.203076 | 1.180575 | 1.182396 | 1.181884 | 1.182179 | 1.182507 | 1.182868 |
| Sparse variogram | 0.391075 | 0.388462 | 0.388475 | 0.389029 | 0.389132 | 0.389351 | 0.389395 |
| Quantile risk | 0.594202 | 0.569788 | 0.569299 | 0.570942 | 0.571087 | 0.571342 | 0.571527 |
| Coverage error | 0.203922 | 0.110632 | 0.107886 | 0.109786 | 0.109589 | 0.109550 | 0.108967 |
| Tail error | 0.129076 | 0.077455 | 0.076718 | 0.078108 | 0.078229 | 0.077396 | 0.077849 |
| Pair variogram | 0.098663 | 0.098663 | 0.098663 | 0.098663 | 0.098663 | 0.098663 | 0.098663 |

Unit labels identify the evaluation representation, not the physical dimension of every metric; normalized feature Energy and calibration errors retain their original definitions in both representations.

**What the action protects.** The action changes member anomalies while protecting the ensemble mean, declared pair differences and orthogonal complement, with a specified low-rank covariance action. It does *not* inherit the cumulative-endpoint protection of the increment compiler. Independent NumPy reconstruction has maximum action discrepancy $7.55 \times 10^{-15}$ and protection residual $5.57 \times 10^{-15}$; all ensembles have eight distinct members and no FIT reference copies. An unchanged registered pair variogram follows from the protected geometry and is not an additional forecast-quality improvement.

**Simple controls retain the Energy advantage.** DD improves month-equal Energy over q0 by 3.3811%/3.3755%, but its loss is 0.0414%/0.0617% higher than scalar and 0.0941%/0.0777% higher than static. DD is only 0.0294%/0.0217% below WW, while DW is below DD. No new confidence intervals or equivalence tests accompany these small differences; numerical proximity is not proof of statistical equivalence. DD's marginal CRPS *increases* by 0.8453%/0.7681%.

All six plugins meet the prior diagnostic allowances (2% relative for five secondary losses and 0.01 absolute for two calibration errors) in both units and aggregates. Meeting a tolerance does not mean every metric improves, nor imply physical safety. The complete primary metric table retains this adverse result.

**Decision and remaining limits.** The frozen stopping rule applies: do not expand seeds, tasks or tuning on this evaluation panel merely because DD did not beat simpler controls. The result establishes descriptive calibration headroom on a fixed stronger backbone, not the necessity of conditioning, a unique executor, or SOTA. There is one fitted instance per head; sampling tapes are not independent training fits. The finite coupled-ensemble score does not inherit an iid unbiased population-risk interpretation automatically. All controlled samples were persisted before this evaluator loaded targets, but previous 2024 exposure remains. Independent terminal checks validate saved actions, scores and aggregates, not complete historical non-access or backbone optimality. An additional experiment is justified only by an independently meaningful request and a declared cost of damaging its protected quantities, compared against simple methods with the same information, constraints and refusal option. Projection identity or successful numerical certification alone does not supply that missing application value.

## S BLOCK-LAW PROTECTION AND A SAME-SUMMARY MECHANISM CHECK

**A separate protection contract.** The affine plugins above can change marginal predictive laws. This exploratory branch tests control that preserves each block's empirical path law. It partitions the 20 stations into ten declared pairs and permutes whole $H \times 2$ member blocks, with $H = 16$ and $M = 8$. For block $b$ and a permutation $\pi_b$, define $X'_{m,b} = X_{\pi_b(m),b}$. Then

$$\frac{1}{M}\sum_{m=1}^{M} \delta_{X'_{m,b}} = \frac{1}{M}\sum_{m=1}^{M} \delta_{X_{m,b}}.$$

Consequently, the multiset of any block-local function is unchanged. This protects the issued local path distribution, including its errors. It does not make that distribution calibrated. Cross-block events and individual member identities are not protected. In particular, this does not inherit the original KRPC pathwise pair-difference or protected-complement certificate. Combining low-dimensional forecasts by multivariate reordering already has precedents (Schefzik, 2016); we do not claim block preservation itself as a new method. Our local controls are not an implementation of that published method. Pairs were chosen using FIT coverage and geographic proximity, not a physical conservation law or a demonstrated operational utility.

**Matching the representation isolates the objective comparison.** Gram matching uses history-derived level/slope summaries. An earlier full-trajectory Energy control changed both the objective and its representation. We therefore add an Energy-distance control using exactly the same 20-dimensional, member-normalized level/slope summaries as Gram matching. All searches start from the same saved ensemble, use past 96-step histories, and share two lexicographic sweeps, 560 swap proposals, at most 20 accepted swaps and a strict $10^{-12}$ improvement threshold. The objectives choose different swaps. Equal search opportunities do not imply equal optimization error or a full representation-by-objective factorial design. No verifying future enters a swap decision. Historical quality exposure is nevertheless retained in the evidence classification.

**Chronology and reuse.** A preliminary screen used 117 consumed 2023 FIT origins. Gram, full-trajectory Energy and same-summary Energy changed standardized finite-measure Energy relative to the base by $+0.1681\%$, $-0.3240\%$ and $+0.0665\%$ (positive means lower loss). These are in-sample mechanism diagnostics. The subsequent fixed comparison uses the same 106 of 347 available 2024 origins and three sampling tapes for each of two fitted backbones. It reuses 1,908 saved

base/Gram/full-trajectory outputs and scores; only 636 same-summary reorderings are new, for 2,544 comparison ensembles in total. There is no new training, base-model sampling or hyperparameter search. The new deterministic stage took 19.12 seconds on CPU; this excludes the earlier training, sampling and independent audit costs. CSDI was adapted to the new panel, whereas D3U transfers an older fitted backbone by the frozen re-panel mapping. Within-backbone controls are matched; absolute scores are not a compute-matched architecture ranking.

**Scores and inferential units.** The primary endpoint is Energy of the issued finite measure, including diagonal terms with denominator $2M^2$. The off-diagonal U-score is retained as a historical sensitivity, not asserted unbiased for these coupled members. We average tapes within origin, origins within month and then the 12 months equally. Month counts are $4, 3, 8, 18, 5, 22, 10, 4, 12, 7, 5, 8$. Paired circular month-block bootstrap uses 20,000 saved draws for each block length 1, 2 and 3; length 2 is primary. With only 12 seasonal months, overlapping windows, missingness selection and repeated panel exposure, these intervals are descriptive, not fresh confirmatory coverage guarantees. Three tapes are not training seeds, and two backbones on one panel are not two independent datasets.

Table 57: Block-law comparison: relative Energy reductions (%) against each backbone's own base, month-equal. Higher is better; negative values are harm. The base is zero by definition. V scores the finite measure; U is the off-diagonal sensitivity. No arm is selected retrospectively as the candidate.

| | | V Energy | | U Energy | |
|---|---|---|---|---|---|
| Backbone | Objective | Standardized | hPa | Standardized | hPa |
| CSDI | Gram | 0.3130 | 0.3375 | 0.3717 | 0.3967 |
| | Full-trajectory Energy | -0.1419 | -0.0492 | -0.2368 | -0.1366 |
| | Same-summary Energy | 0.2282 | 0.2676 | 0.2438 | 0.2852 |
| D3U | Gram | 0.1111 | 0.1256 | 0.1263 | 0.1400 |
| | Full-trajectory Energy | 0.1428 | 0.1734 | 0.1475 | 0.1825 |
| | Same-summary Energy | 0.1039 | 0.1096 | 0.1047 | 0.1078 |

Table 58: Gram minus same-summary Energy: absolute V-score differences, negative favoring Gram. Primary length-2, 95% descriptive percentile intervals all cross zero, as do lengths 1 and 3. This establishes neither superiority nor equivalence. The last column counts months with lower Gram loss.

| Backbone | Unit | Mean difference | Interval | Months |
|---|---|---|---|---|
| CSDI | Standardized | -0.0002137 | [-0.0005450, 0.0001054] | 8/12 |
| | hPa | -0.0013972 | [-0.0040196, 0.0013455] | 8/12 |
| D3U | Standardized | -0.0000196 | [-0.0002845, 0.0002324] | 5/12 |
| | hPa | -0.0003433 | [-0.0021265, 0.0013811] | 7/12 |

**Attribution and counterexample.** Tables 57–58 narrow the interpretation of the earlier CSDI result. Holding the summary fixed reduces Gram's mean advantage over its Energy control to $0.0851\%/0.0701\%$ (standardized/hPa), compared with $0.4543\%/0.3865\%$ against full-trajectory Energy. Same-summary Energy beats full-trajectory Energy in all 12 CSDI months, with all three block-length intervals below zero. On D3U, Gram's same-summary advantage is only $0.0072\%/0.0160\%$, while full-trajectory Energy has the best mean of these three objectives. Thus the predeclared both-backbone, all-control point-superiority condition fails. The narrower both-backbone, same-summary point condition passes, but does not establish statistical superiority or an irreplaceable Gram objective. Origin-equal sensitivity retains this ordering: Gram improves over the same-summary control by $0.1365\%/0.1207\%$ for CSDI and $0.0065\%/0.0117\%$ for D3U, with full-trajectory Energy still better on D3U.

**Adverse and invariant endpoints.** The CSDI Gram arm worsens sparse variogram by $0.1408\%/0.3755\%$ despite its V-Energy gain. Same-summary Energy has better temporal Energy on CSDI and better month-equal sparse variogram on D3U. These are not all-metric no-harm results. Marginal CRPS, quantile risk, coverage error, tail calibration error and the registered within-pair

variogram are invariant to floating precision, not improved. The largest per-ensemble CRPS difference is $1.12 \times 10^{-16}$ after rounding upward. The other varying endpoints and all three objectives are reported in Table 59.

Table 59: Remaining varying endpoints: relative reductions (%) from each base, standardized/hPa. Temporal Energy is invariant to the common unit change here. Together with Table 57 and the five invariant endpoints, this covers all ten registered scores.

| Backbone | Objective | Temporal Energy | DCT CRPS | Sparse variogram |
|---|---|---|---|---|
| CSDI | Gram | 0.1739 | 0.2392 / 0.1246 | -0.1408 / -0.3755 |
| | Full-trajectory Energy | 0.0417 | -0.0277 / -0.0664 | -1.0086 / -0.8488 |
| | Same-summary Energy | 0.2021 | 0.2087 / 0.1027 | -0.4409 / -0.4087 |
| D3U | Gram | 0.0293 | 0.1623 / 0.4782 | 0.2676 / 0.3108 |
| | Full-trajectory Energy | -0.0020 | 0.0615 / 0.4189 | 0.1415 / 0.3681 |
| | Same-summary Energy | -0.0030 | 0.0654 / 0.2936 | 0.3765 / 0.4014 |

**Audit scope and stopping decision.** The completed audit verifies 1,337 source files, a continuous 1,279-row ledger, exact reuse of old control samples and scores, block laws, eight distinct paths and zero reference copies. Independent Energy/CRPS formulas agree within $4.45 \times 10^{-15}$, and weights and bootstrap reductions are checked independently. Real optimizer replay uses the production implementation; independent naive optimizer checks cover synthetic cases only. Historical D3U target isolation and full GPU sample regeneration are not established by this audit. These checks establish internal consistency, not novelty or external validity. We close this fixed mechanism comparison rather than retune on the same outcomes. Further claims would require a justified external request or decision utility, appropriate published comparators and separately designed evaluation. The present evidence is post-hoc mechanism attribution, not fresh confirmation, qualification or field-wide SOTA.

## T SOLAR: DEPENDENCE CORRECTION VERSUS STRONG SIMPLE CONTROLS

**Joint path generation with fixed channel-path marginals.** This bounded hourly Solar comparison uses all $K = 137$ channels, horizon $H = 24$ and $M = 100$ forecast paths. It reuses nine validation-selected checkpoints: TimePrism-product, TimeGrad-native and TACTiS-2-native, each with technical seeds 3141, 3142 and 3143. Seven consecutive daily origins $(7009, 7033, 7057, 7081, 7105, 7129, 7153)$ yield 63 base ensembles. The six fixed arms are q0, history-rank, history-assignment, request-swap, feature-energy-swap and history-energy-swap. Each base ensemble is sampled once and reused by every control. Controls permute entire 24-hour paths within each channel. This changes cross-channel dependence while preserving empirical channel-path marginals. It does not inherit the earlier memberwise KRPC protection contract. Swap arms retain their fixed two sweeps and $10^{-12}$ strict-improvement threshold. No additional training or hyperparameter search uses this test.

The reporting protocol was frozen after complete validation and before this branch opened test values; it is not a validation-blind preregistration. Validation fixed history-rank as TimePrism's strong simple reference and history-assignment for TimeGrad and TACTiS-2. All three swap arms are reported against both q0 and that fixed reference. No swap arm was promoted as a validation winner, and the reference is not reselected on test.

Each origin's scores are averaged within seed, then the three seed means are weighted equally. Energy divided by $\sqrt{HK}$ is primary; empirical marginal and sum CRPS and normalized 19-quantile marginal and sum CRPS are secondary. Empirical scores retain the finite-measure $M^2$ denominator; quantile scores use the frozen per-origin normalization before averaging.

**Simple controls improve q0 more than the swap arms.** Table 60 retains every arm and metric. The fixed simple references reduce Energy relative to their own q0 by 12.5614%, 0.2903% and 0.7363% for TP, TG and TC, respectively. Request-swap reduces it by 8.1909%, 0.0936% and 0.2621%, yet has 4.9984%, 0.1974% and 0.4778% higher Energy than those simple references. Feature-energy-swap improves q0 Energy by 1.4697%, 0.0838% and 0.1181%; history-energy-swap

Table 60: Complete Solar standard-test means in fixed protocol order, not a ranking. All columns are lower-is-better. E is Energy/$\sqrt{HK}$, C and S are empirical marginal and sum CRPS, and Q and QS are normalized 19-quantile marginal and sum CRPS. TP, TG and TC denote the three backbones above. R/A are history-rank/assignment; Req/F/H are request-, feature-energy- and history-energy-swap. Seven origins are averaged within each of three technical seeds, then seeds equally. Both marginal columns are exactly unchanged before rounding. No confidence interval or independence claim is attached to these descriptive means.

| Base | Arm | E | C | S | Q | QS |
|---|---|---|---|---|---|---|
| TP | q0 | 27.671549 | 12.382873 | 2053.836253 | 0.498822 | 0.644097 |
| TP | R | 24.195618 | 12.382873 | 1540.966148 | 0.498822 | 0.474533 |
| TP | A | 24.211875 | 12.382873 | 1486.648602 | 0.498822 | 0.467506 |
| TP | Req | 25.405013 | 12.382873 | 1894.562440 | 0.498822 | 0.584834 |
| TP | F | 27.264852 | 12.382873 | 2045.058020 | 0.498822 | 0.643249 |
| TP | H | 25.588576 | 12.382873 | 1910.109365 | 0.498822 | 0.593232 |
| TG | q0 | 22.955452 | 11.730422 | 1533.582278 | 0.518816 | 0.526762 |
| TG | R | 22.928201 | 11.730422 | 1560.676266 | 0.518816 | 0.534072 |
| TG | A | 22.888801 | 11.730422 | 1492.449770 | 0.518816 | 0.515971 |
| TG | Req | 22.933975 | 11.730422 | 1523.210683 | 0.518816 | 0.523327 |
| TG | F | 22.936205 | 11.730422 | 1521.696760 | 0.518816 | 0.524726 |
| TG | H | 22.958473 | 11.730422 | 1537.597925 | 0.518816 | 0.527738 |
| TC | q0 | 21.756202 | 11.178085 | 1430.022691 | 0.458726 | 0.455706 |
| TC | R | 21.685877 | 11.178085 | 1394.174016 | 0.458726 | 0.447291 |
| TC | A | 21.596002 | 11.178085 | 1304.019288 | 0.458726 | 0.422788 |
| TC | Req | 21.699181 | 11.178085 | 1399.522408 | 0.458726 | 0.448760 |
| TC | F | 21.730500 | 11.178085 | 1410.726432 | 0.458726 | 0.455079 |
| TC | H | 21.739097 | 11.178085 | 1421.733380 | 0.458726 | 0.454382 |

improves TP and TC by 7.5275% and 0.0786%, but worsens TG by 0.0132%. All three swap arms have worse mean Energy and both sum-CRPS measures than the corresponding fixed simple reference on each backbone. The marginal scores are identical by construction, not an additional improvement. These observations do not establish statistical equivalence, significance, or superiority of the swap mechanism.

**Adverse outcomes and heterogeneity remain visible.** For TG, history-energy-swap also worsens empirical sum CRPS by 0.2618% and quantile sum CRPS by 0.1853% against q0. History-rank improves TG Energy by 0.1187% but worsens these aggregate scores by 1.7667% and 1.3878%. Even an improving mean need not describe every technical seed: TC feature-energy-swap increases seed-3142 Energy from 21.153338 to 21.177593, although its three-seed mean improves. The companion complete results record retains all 54 seed rows, 24 absolute and 24 relative fixed contrasts, and 168 origin-paired contrasts alongside the 18 headline means; it does not discard negative contrasts or select favorable days.

**Provenance and limits of the conclusion.** Every base and controlled ensemble was committed before the evaluator persisted the first test target. Independent terminal checks cover all 63 same-address neural replays, all control actions, five-metric arithmetic and equal-seed aggregation. Saved neural replays match the original bases with maximum absolute error zero under the original $10^{-4}/10^{-5}$ absolute/relative tolerance, and every arm retains 100 distinct paths. Source and input bytes, past-only boundaries, resource receipts and RNG evidence were checked; operational interruptions and their original records were retained. Replay uses the same neural implementation, not an independent derivation of the model mathematics. Audit success is therefore an implementation and evidence-chain result, not a performance theorem.

Seven consecutive days do not become 21 independent test days by using three technical seeds; no new confidence interval, significance test or temporal-independence assumption is introduced. Controllers use 2400 past hours, versus 192 for TP/TG and 24 for TC, so the comparison is not an equal-information ablation. Later rolling origins may use earlier observed test-period history without retraining. TC's stage-two validation-best is at the fixed-budget endpoint, which does not prove

Table 61: Adaptive Solar validation means for the fixed simple start (S) and all three matched residuals: Gram (G), feature-Energy (F) and full-Energy (H). All five columns are lower-is-better, with definitions as in Table 60. Ten origins are averaged within each technical seed, then the three seeds equally. Original six-arm, per-seed and fixed per-origin contrasts remain in the complete results record. This table is not a fresh standard-test or SOTA comparison.

| Base | Arm | E | C | S | Q | QS |
|---|---|---|---|---|---|---|
| TP | S | 24.294061 | 12.248956 | 1484.784113 | 0.705862 | 0.704984 |
| TP | G | 23.930674 | 12.248956 | 1473.992851 | 0.705862 | 0.696583 |
| TP | F | 27.537554 | 12.248956 | 1705.824882 | 0.705862 | 0.809501 |
| TP | H | 25.423043 | 12.248956 | 1567.266878 | 0.705862 | 0.748929 |
| TG | S | 20.644172 | 10.875684 | 1281.086704 | 0.609365 | 0.591515 |
| TG | G | 20.645842 | 10.875684 | 1277.345612 | 0.609365 | 0.589457 |
| TG | F | 20.642099 | 10.875684 | 1286.483952 | 0.609365 | 0.593162 |
| TG | H | 20.649244 | 10.875684 | 1283.661771 | 0.609365 | 0.592320 |
| TC | S | 20.899861 | 10.694910 | 1248.076118 | 0.623526 | 0.592068 |
| TC | G | 20.890183 | 10.694910 | 1239.065974 | 0.623526 | 0.588548 |
| TC | F | 20.923789 | 10.694910 | 1260.579243 | 0.623526 | 0.596543 |
| TC | H | 20.905280 | 10.694910 | 1249.739088 | 0.623526 | 0.592887 |

optimization saturation. This bounded hourly setting does not reproduce the paper-best Solar-10min setting or exhaust contemporary methods. The results support descriptive dependence-correction headroom, often captured better by simple controls; they do not support an indispensable swap mechanism or field-wide SOTA. No further tuning or experiment selection on these consumed test outcomes is part of this study.

## U SOLAR: STRONG-START RESIDUAL CONTROL AS ADAPTIVE DEVELOPMENT

**Residual control from a strong simple permutation.** After observing the original validation and standard-test results, we tested whether two-sweep control could improve a strong simple permutation. This is adaptive validation development, not new confirmation. We reused the original 90 validation base ensembles and their 2400-hour past histories: three backbones, three technical seeds and ten daily origins, with $M = 100, H = 24, K = 137$. TimePrism starts from history-rank; TimeGrad and TACTiS-2 start from history-assignment, retaining the references fixed before the original test. Gram, feature-Energy and full-Energy residuals use the same start, two sweeps, candidate enumeration and $10^{-12}$ improvement threshold. Residual permutations are composed with the start and certified against the original base. All original six arms are retained in the complete record. All 270 new actions were saved before new target access or scoring. No neural training, forecast sampling or standard-test reevaluation was performed.

**A useful but backbone-dependent development signal.** Against its fixed simple start, the Gram residual improves TimePrism Energy by 1.4958%, empirical sum CRPS by 0.7268% and quantile sum CRPS by 1.1917%. For TACTiS-2 the improvements are 0.0463%, 0.7219% and 0.5946%, respectively. On these two backbones it also beats both matched Energy residuals on all three non-invariant mean scores. TimeGrad instead has 0.0081% worse Energy than the simple start and 0.0181% worse Energy than the feature-Energy residual, although its two sum scores improve 0.2920% and 0.3479% over the start. The Energy controls also show adverse outcomes: both worsen all three non-invariant means for TP and TC; TG feature-Energy improves Energy but worsens both sum scores, and TG full-Energy worsens all three. Marginal CRPS remains exactly invariant.

**The global investment rule fails on one backbone.** Before this residual run, we fixed a descriptive investment rule requiring Gram Energy to beat the simple start and both matched residuals on every backbone, without any secondary-mean increase. This rule fails because of TimeGrad; we do not replace it with a two-of-three criterion or expand the search budget automatically. These development observations support backbone-dependent value from the strong-start Gram residual, not

universal indispensability, statistical significance or field-wide SOTA. Ten daily origins with three technical seeds are not thirty independent days, and controls retain their longer history than the original model contexts. The original standard-test negatives are unchanged. Terminal checks cover 90 units, 270 exact persisted action replays, finite five-metric arithmetic and equal-seed aggregation; every new ensemble has 100 distinct paths. The optimizer implementation is reused for deterministic replay, while composition and metric arithmetic are checked independently. This audit does not turn adaptive development into independent quality confirmation.

# V Extended evidence and disclosure boundary

Five labels remain distinct: known-truth qualification, deterministic certificate, prospectively frozen confirmation, target-informed development, and implementation-only history. The K32/K128 atlas has the first label; KRPC residuals the second; the frozen EEG confirmation (D25 experiment ID) and separately frozen USGS outer roster the third; PTB-XL folds 1–7 the fourth; and the interrupted fold-8 execution the fifth. "Frozen" denotes internal pre-specification before the 61 confirmation targets were opened, not an external preregistration. The USGS label applies only to zero-request roster promotion; it does not complete an independent second-family test of the proposed generator–control conjunction.

The legacy diffusion-paper directory contains Electricity-321, Traffic-862, Weather, Exchange, ETT, Solar, financial/crypto files, and the frozen GHCNh station panel. Their availability does not establish fresh confirmation. The initial GHCNh 2022–2023 past-only observability study remains unresolved; the separately frozen repair subsequently used 2024 past windows for mechanism evaluation, as reported in Appendix P. It did not score a future target, and did not access 2025 in that mechanism study. A separate completed project evaluation subsequently scored the old stations' 2025 targets; QC-only exposure is therefore not a sufficient description of that year. The later nearby-station evaluation (Appendix P.9) also scores 2025 but is limited to 55/346 eligible origins. Neither source checks nor geographic separation establish global freshness or independence. Earlier GHCNh quality and orientation diagnostics reuse consumed 2022–2023 outcomes. The subsequently authorized re-panel scores already exposed 2024 targets, including the matched CSDI and plugin comparisons in Appendix R; sample-before-score execution does not erase that project history. MERRA-2 2020 remains adaptive development. Its separately frozen 2022 selection and locked 2023 confirmation are now complete (Appendix M); 2021 is historical development, not an unused holdout. After that development and the old 2023 result, the multiscale correction was locked and tested on 2024 (Appendix N). This revision incorporates that authorized result. The subsequent contextual expert gate was frozen on 2020 development and evaluated once on 2025 (Appendix O); no MERRA-2 annual holdout remains fresh for another method iteration. Unread PTB-XL folds remain unopened.

AI assistance covered hypothesis refinement, methodological critique, experimental planning, implementation and debugging support, literature organization, figure planning, and manuscript drafting and revision. Human authors retain responsibility for source verification, target-access decisions, statistical interpretation, claim authorization, and the final submission.

# W GPM IMERG: NEGATIVE SELECTION AND ADAPTIVE FIT BOUNDARY

The GPM IMERG extension tests whether the method passes selection in another data family. It failed entry and provides no second-family positive confirmation. Its 2020 development completed 4,224 origin–unit evaluations across 22 execution units. The 2021 selection then completed 1,728 evaluations: 192 origins, three equally weighted replicates, and three arms (positive-affine Q1, an unprotected proposal, and the protected candidate), with $K = 128$, $H = 16$, and $M = 64$. Within each replicate–origin address, all arms reused the same persisted Q0 ensemble. The primary was full-path fair Energy; point MAE did not select the model.

Relative to Q1, the candidate reduced fair Energy by 0.013550%, marginal fair CRPS by 0.008593%, increment fair Energy by 0.010258%, path-q99 CRPS by 0.109384%, pooled-q99 absolute error by 0.013104%, and soft-event Brier by 0.011473%. Spatial variogram instead increased by 0.008271%: its ratio was 1.0000827078 against the frozen no-harm limit of 1.0. That single failure rejected entry. These are descriptive selection differences, not significance claims. All candidate hard-event and stated mean, paired-channel-difference and channel pooled-q99 certificates passed, along with 64-distinct and zero-copy checks against the registered fit-only reference bank. Those controls do not protect every cross-channel variogram term; neither the small violation nor correct execution permits a post-hoc tolerance or a qualification claim.

**Execution amendment.** This was not an uninterrupted, unamended one-shot. Execution stopped after 1,622 of 1,728 evaluations, before candidate replicate 2 origin 086, at a pair-shared increment guard. An authorized post-unsealing amendment retained the old allclose acceptance and additionally admitted inputs only when the original shared residual construction matched bitwise. The resumed tail reused existing bases and samples; direction, projection, final certificate tolerances, seeds, $M$, scores and entry gates were unchanged. A separate scheduling amendment allowed the last 106 CPU-only origins to coexist with external MPS work under an 8 GiB available-memory start gate. This disclosure limits claims about execution pre-specification; the failed final entry decision is retained rather than relabeled as a pristine positive test.

**Adaptive development after selection.** Subsequent restoration studies reused 36 already-observed 2017–2019 fit addresses (12 per replicate). Across the post-projection, full-support, scale-aware, feasibility-gate, feasible-direction and stable-arithmetic branches, at least 154 control-layer updates were performed, in addition to upstream fitting, diagnostics and audits. The matched direction branch completed 24 updates and 83 candidate trials but qualified none of three replicates; mean replicate violation was 3.098057% worse than the preceding feasibility-gate branch. Its stable-arithmetic follow-up completed another 24 updates and 83 trials, again qualifying none of three replicates. It preserved the accepted step sequence and did not materially repair quality; its implementation audit is not positive generalization evidence. A post-hoc analysis of all 24 saved gradients found that earlier increment protection changes nine directions in replicate 2 while retaining first-order common descent; it changes neither other replicate and has not been validated by new generated paths. First-order feasibility, plateau stopping and implementation audits therefore establish neither practical recovery nor local optimality. These adaptive fit results are not independent replication. GPM 2021 is consumed selection data; qualification 2022 and outer 2023–2024 remain unopened. The GPM result does not change the distinct MERRA-2 year-access history.

# X STRUCTURAL ENDPOINTS: DISTINGUISH LEGITIMATE VARIATION FROM FAILURE

This study tests the evaluator, not a learned generator. Its protocol was fixed before the calibration and challenge draws. It complements the frame and request qualification in Appendix D; it does not retroactively qualify every endpoint used elsewhere in the paper.

## X.1 PROCESSES, CORRUPTIONS AND CALIBRATION

Each independent trial contains a reference ensemble, a candidate ensemble and a bank of 16 paths, drawn independently. Ensembles have $M = 64, H = 32, K = 32$. Two phases contain 16 time steps each. Independent unit-variance latent processes $u, v$ define $X_1 = u$ and

$X_2 = \rho_t u + \sqrt{1-\rho_t^2}v$, with $\rho_t = 0.6$ in the first phase and $-0.6$ in the second. The remaining channels are independent processes of the same family. The three families are Gaussian iid, Gaussian AR(1) with coefficient 0.6, and AR(1) with standardized $t_5$ innovations. Gaussian initial states are stationary; the heavy-tailed process starts at zero and uses a 128-step burn-in. The mixed second channel need not itself have a $t_5$ marginal. Matrix diagnostics observe only the first two coordinates; the copy diagnostic observes complete paths. Embedding the task in $K = 32$ does not qualify all $32 \times 32$ dependencies.

Four paired variants use the same trial. A legal candidate is a completely new ensemble with the prescribed phase change. A flattened candidate replaces the second channel by $v$, eliminating the phase relation. In the Gaussian families, its single-time marginals and population pooled covariance are unchanged, although its phase covariance differs by squared Frobenius effect $2(0.6)^2 = 0.72$. This marginal statement need not hold for the heavy-tailed mixture. A within-phase time permutation uses the order $[0, 8, 1, 9, \ldots, 7, 15]$ for all channels and members. It preserves empirical phase covariance, is law-preserving in the iid family, and disrupts the AR families' lag structure. Finally, near-copy candidates cycle through the 16 bank paths and add independent Gaussian noise with standard deviation 0.01. Exact distinctness alone need not detect these noisy copies.

The phase statistic averages the squared Frobenius difference between candidate and reference phase covariance matrices (population divisor). The lag statistic instead compares their phase-specific lag-one matrices, using pooled member–time centering within each phase and no cross-boundary pairs. The near-copy statistic is the negative mean nearest-bank full-path RMSE. Larger values are more suspicious for all three. Full-block pooled covariance is a diagnostic comparator, not part of the rejection rule. There is no time warping or samplewise normalization to remove the tested differences.

For each family, 1,024 legal trials calibrate the statistics. Each threshold is the maximum across families of their 1,022nd ordered calibration value, where $1022 = \lceil (1024+1)(1-0.01/3) \rceil$. Exceedance is strict; ties do not reject. Thresholds are saved before any challenge trial. The omnibus rule rejects when any of the first three statistics exceeds its threshold. Challenge uses 512 new independent trials per family, with four paired variants per trial: 6,144 rows, not 6,144 independent trials. NumPy PCG64 addresses use SeedSequence $[2609268201, s, f, i, j]$, where $s = 0, 1, 9$ denotes calibration, challenge or engineering tests, and $j = 0, 1, 2, 3$ denotes reference, candidate, bank or copy perturbation. These addresses do not reuse the earlier frame-qualification tape.

### X.2 All primary assertions and their limits

The 12 predeclared assertions comprise four legal-control false-positive rates and eight designated-endpoint detection rates. Each uses an exact two-sided Clopper–Pearson interval at confidence $1 - 0.05/12$, giving at least 95% simultaneous coverage by Bonferroni. Qualification requires every false-positive upper bound to be at most 5% and every detection lower bound to be at least 80%. Trials are independent conditional on the saved thresholds; neither members nor time steps increase the inferential sample size. Pairing does not require independence between the 12 assertions.

The two Gaussian phase-flattening challenges trigger the phase test in 512/512 trials each, but the pooled test in 0/512. The heavy-tailed pooled count is 3/512, so the diagnostic is not identically blind. Time permutation triggers no legal iid rejection, yet the designated lag test detects 512/512 in each AR family. The three endpoints are complementary, not a complete characterization of the joint law. In particular, phase covariance ignores within-phase order, and the copy test covers a finite bank and one noise scale, not privacy or all memorization attacks.

Calibration and challenge share the same three DGP families. Independent random addresses give sample independence, not unseen-family transfer. The correlation, AR coefficient, perturbation amplitude and observation budget are fixed effects; power for subtler failures remains unmeasured. Nor does the pooled-covariance example imply that the full joint Energy score is theoretically blind. Independent arithmetic, threshold selection, all addresses and simultaneous intervals were verified after the single 29.71-second two-thread CPU run. Only statistics were persisted, not all synthetic paths. This is endpoint evidence at the declared scope, not a generator leaderboard or an estimated power guarantee for Solar or EEG.

Table 62: **All 12 structural-endpoint assertions.** Each count has denominator 512. Intervals in percent have simultaneous coverage at least 95%; they are not separate unadjusted 95% intervals. FP denotes a legal control's omnibus false-positive rate. Det. denotes the specified endpoint's detection rate. All assertions meet the fixed bounds; nonzero heavy-tail false positives remain visible.

| Family | Variant | Endpoint | Role | Count | Simultaneous interval, % |
|---|---|---|---|---|---|
| Gaussian iid | New | Omnibus | FP | 0/512 | [0.000, 1.199] |
| Gaussian iid | Flatten | Phase | Det. | 512/512 | [98.801, 100.000] |
| Gaussian iid | Permute | Omnibus | FP | 0/512 | [0.000, 1.199] |
| Gaussian iid | Near-copy | Copy | Det. | 512/512 | [98.801, 100.000] |
| Gaussian AR | New | Omnibus | FP | 0/512 | [0.000, 1.199] |
| Gaussian AR | Flatten | Phase | Det. | 512/512 | [98.801, 100.000] |
| Gaussian AR | Permute | Lag | Det. | 512/512 | [98.801, 100.000] |
| Gaussian AR | Near-copy | Copy | Det. | 512/512 | [98.801, 100.000] |
| Student AR | New | Omnibus | FP | 7/512 | [0.340, 3.575] |
| Student AR | Flatten | Phase | Det. | 512/512 | [98.801, 100.000] |
| Student AR | Permute | Lag | Det. | 512/512 | [98.801, 100.000] |
| Student AR | Near-copy | Copy | Det. | 512/512 | [98.801, 100.000] |

# Y A COMMON REQUEST–PROTECTION TASK ON CACHED SOLAR ENSEMBLES

This post-hoc mechanism study uses only 90 previously generated validation ensembles: three backbones (channel-product TimePrism, native TimeGrad and native TACTiS-2), three technical seeds and ten consecutive origins. Each ensemble has $M = 100$, $H = 24$, $K = 137$. The same base serves every control; neither targets, history arrays nor standard-test payloads enter this task. The existing signed model coordinates support a generic control test, not a claim about physically nonnegative power or grid value. The caches have already been exposed in earlier validation and feasibility studies.

## Y.1 PUBLIC REQUEST, MANDATORY PROTECTION AND SECONDARY COSTS

Using zero-based channel indices, two orthonormal basket columns average even and odd channels with weights $1/\sqrt{69}$ and $1/\sqrt{68}$. These are public index groups, not geographic regions. For each phase $[0, 12)$ or $[12, 24)$, each member is centered over time, giving residuals $R$ and basket coordinates $Z = RU$. The empirical request covariance is $C = Z_{\mathrm{flat}}^{\top} Z_{\mathrm{flat}}/(100 \cdot 12)$. This differs from the reference-law statistics in the preceding appendix: it is a numerical finite-ensemble constraint, not a population-quality test.

The three fixed requests are identity ($T = C$), constant relation, and phase reversal. The latter two use $T = ACA^{\top}$ with $A = R_\theta \operatorname{diag}(1.25, 0.8) R_\theta^{\top}$. Constant relation takes $\theta = \pi/4$ in both phases; phase reversal takes $\pi/4$ then $-\pi/4$. Every method receives the same base, $U$ and $T$, not the defining map $A$. The common domain requires the smallest eigenvalue of $C$ to exceed $10^{-12}$ and its condition number to be at most $10^8$. Out-of-domain requests are refused for all methods.

The seven controls are identity, optimal scalar, diagonal, unrestricted affine, protected whiten–color, dense projected OT and thin KRPC–OT. Scalar uses $s^2 = \max(0, \langle C, T\rangle_F / \langle C, C\rangle_F)$; diagonal uses entries $\sqrt{T_{ii}/C_{ii}}$. Whiten–color uses $T^{1/2} C^{-1/2}$. OT uses the standard symmetric map $C^{-1/2}(C^{1/2} T C^{1/2})^{1/2} C^{-1/2}$. Unrestricted affine applies this map to uncentered basket coordinates; protected methods apply their map to $R$ and add the resulting basket edit back to the original path. Dense OT forms the full channel action; KRPC uses the thin implementation of Equation 3 phasewise. Identical exact-arithmetic OT outputs are expected, not a unique advantage of one executor. KRPC retains its gain envelope $[0.5, 2]$.

The primary endpoint is joint request-and-protection success on all nonidentity requests, including refusals in the denominator. Request relative Frobenius error must be at most $10^{-8}$ in both phases. With $b = \max(1, \max |X|)$, the maximum complement error divided by $b$, and phase-sum error divided by $12b$, must each be at most $10^{-10}$. Every edited ensemble must retain 100 distinct full paths and have no exact copy of any member of its finite base ensemble. This is not a training-set

Table 63: **All controls on the same nonidentity requests.** Entries are joint successes out of 60 per backbone (three technical seeds, ten origins, two requests). These are descriptive counts, not independent-trial confidence statements. All seven current controls remain; the original KRPC numerical implementation is shown separately, not erased. Each current control also passes all 90 identity positive controls.

| Control | TimePrism | TimeGrad | TACTiS-2 |
|---|---|---|---|
| Identity | 0/60 | 0/60 | 0/60 |
| Optimal scalar | 0/60 | 0/60 | 0/60 |
| Diagonal | 0/60 | 0/60 | 0/60 |
| Unrestricted affine | 0/60 | 0/60 | 0/60 |
| Protected whiten–color | 60/60 | 60/60 | 60/60 |
| Dense projected OT | 60/60 | 60/60 | 60/60 |
| KRPC–OT, common-unit v2 | 60/60 | 60/60 | 60/60 |
| KRPC–OT, original v1 (retained) | 9/60 | 18/60 | 11/60 |

or near-copy guarantee. Identity has its own 90 positive controls; unchanged identity paths are not incorrectly rejected as copies.

Secondary outcomes measure changes relative to the base: relative Frobenius edit, marginal empirical W1 divided by mean absolute base magnitude, relative phase-specific basket lag-one matrix change, and the change in the fraction of negative entries. They are reported costs, not additional post-hoc vetoes. No request-matched true law is supplied, so these changes are not labeled improved or degraded generation quality. Equal averages over technical seeds, origins and requests are descriptive; ten consecutive origins and three seeds do not provide 30 independent real-data trials.

Identity, scalar and diagonal preserve the named structure but miss the nonidentity requests. Unrestricted affine realizes the covariance request but breaks the phase sums. The three protected full-matrix controls attain the best observed primary success rate together. They demonstrate the capability combination, not its exclusivity. All 1,620 original comparator rows are reused unchanged; only the 270 KRPC rows, including identity, are recomputed in the numerical revision.

### Y.2 The numerical failure, repair and retained trade-offs

The original KRPC implementation succeeds on 38/180 nonidentity requests and refuses 142. Every first failing phase rejects a mathematically symmetric intermediate matrix $C^{1/2}TC^{1/2}$ under an absolute $10^{-12}$ symmetry test. Its magnitude ranges from $5.72 \times 10^{7}$ to $5.41 \times 10^{11}$; relative asymmetry is only $1.05 \times 10^{-16}$–$3.28 \times 10^{-16}$. The dense comparator removes roundoff asymmetry before decomposition. This implementation discrepancy is a real failure of v1, not insufficient training or a secondary-metric cost.

A separately specified v2 divides both $C$ and $T$ by their common maximum absolute entry, accepts only relative asymmetry at most $64\epsilon_{64}$, and removes that rounding component before calling the unchanged OT core. The identity $G(C/s, T/s) = G(C, T)$ leaves the control problem unchanged. The original-unit residual, common domain, gain and protection thresholds remain in force. The entire task is not invariant to arbitrary units, because its original absolute minimum-eigenvalue condition still applies. Before its single run, 66 synthetic numerical tests cover multiple ranks, condition numbers and scales $10^{-18}$–$10^{18}$, while retaining true asymmetry, non-SPD and gain violations as refusals. Those tests are not additional real-data efficacy trials.

After repair, all 180 edits succeed with zero refusals. Independent reconstruction verifies every new output and primary/secondary arithmetic. The maximum scaled thin/dense output difference on mutually applied requests is $3.53 \times 10^{-12}$. This comparison does not mix a refused base fallback with an executed dense action. The 6.88-second two-thread CPU revision reuses existing bases and comparator scores; it is not a controlled runtime benchmark against other methods.

The independent audit also verifies all 270 request commits, the full 1,890-row seven-arm aggregation and unchanged comparator records. Current source texts and input/parent file identities remain consistent. File identity checks are not independent historical byte authentication since the earlier

Table 64: **KRPC v2 retains visible secondary costs.** Nonidentity averages, relative to the original signed base. W1 and lag change are dimensionless; negative-entry change is in percentage points. These are not errors against a verifying future. No physical nonnegativity claim is made.

| Backbone | Relative edit, % | W1/base | Lag-one change | Negative change, pp |
|---|---|---|---|---|
| TimePrism | 12.4208 | 0.239526 | 0.511500 | +10.2268 |
| TimeGrad | 15.4286 | 0.250223 | 0.511556 | +9.4572 |
| TACTiS-2 | 15.5630 | 0.262667 | 0.511355 | +16.6717 |

neural audit; that audit is reused, not rerun. Full control and secondary records retain both versions, all methods and all seeds.

This experiment supports execution of a specified finite request family with explicit protection. It does not establish unknown-regime adaptation, downstream utility, nonnegative generation, or a generation-quality SOTA. The synthetic challenge above and this cached-path task answer different questions; their successes cannot be combined into an untested guarantee that every required capability holds on one unknown real process. The fixed task ends here, without searching the exposed caches for requests that would turn a tie with strong controls into an exclusive lead.

# Z TASK-MATCHED EVALUATOR COMPARISON ON INDEPENDENT DRAWS

This study asks whether declared dynamic failures can be detected without rejecting legitimate new trajectories. It compares eight evaluation statistics on the same samples, not eight generative models. The protocol was fixed before its new calibration and challenge draws. It does not replace the earlier qualification in Appendix X: the observation budget, centering convention and random namespace are different. Energy is not assumed to require stationarity. Its finite-sample detection power is the quantity tested here, not its population identification property (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015).

## Z.1 PROCESSES, LEGAL CONTROLS AND FIXED FAILURES

Each trial contains independent candidate and reference ensembles with $M = 32, H = 32, K = 16$. The three declared families are Gaussian iid, Gaussian AR(1) with coefficient $0.6$, and AR(1) with coefficient $0.6$ and innovations $\sqrt{3/5}\, t_5$. Gaussian initial states are stationary standard normals. The Student process starts at zero and uses 128 burn-in steps. For independent latent channel processes $u, v$, the two public channels are $X_1 = u$ and $X_2 = \rho_t u + \sqrt{1-\rho_t^2}v$, with $\rho_t = 0.6$ for the first 16 steps and $-0.6$ for the last 16. The other 14 channels are independent nuisance processes from the same family. The phase boundaries and two public channels are given to every task-matched evaluator; they are not selected from challenge outcomes.

Each family has 1,024 legal calibration trials and 512 independent challenge trials. Five paired candidate variants share each challenge reference. *New* is an unchanged, independently drawn candidate ensemble. *Half* recovers its latent $v$ and replaces $\rho_t$ by $\rho_t/2$; *flat* sets that correlation to zero. Both use $\sqrt{1-(\rho_t^{\mathrm{new}})^2}$ for the complementary coefficient. These transformations preserve Gaussian single-time marginals, but need not preserve the full Student-mixture marginals. *Scramble* applies the same order $[0, 8, 1, 9, \ldots, 7, 15]$ within each phase to all members and channels. This is legal under iid sampling but changes the AR path law. *Shift* adds $0.35$ to the complete candidate. This final variant is a predeclared coverage counterexample, not part of the primary dynamic-power average. No variant consumes new random draws.

NumPy PCG64 addresses use SeedSequence $[2609268601, s, f, i, j]$, with $s = 0, 1, 9$ for calibration, challenge and engineering, respectively; $f = 0, 1, 2$ for family; and $j = 0, 1$ for reference and candidate. Thus the study contains 3,072 calibration and 1,536 challenge trials, producing 7,680 challenge rows. A whole ensemble trial is the independent unit. Its five variants, members and time steps are repeated measurements. New addresses supply sample independence, not transfer to an unseen process family.

### Z.2 Eight statistics with common inputs

For feature vectors $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ of dimension $d$, define the normalized two-sample Energy U statistic

$$\widehat{D}_{\mathrm{E}}(a,b) = \frac{1}{\sqrt{d}} \left[ \frac{2}{mn} \sum_{i,j} \|a_i - b_j\|_2 - \frac{\sum_{i \neq i'} \|a_i - a_{i'}\|_2}{m(m-1)} - \frac{\sum_{j \neq j'} \|b_j - b_{j'}\|_2}{n(n-1)} \right]. \tag{33}$$

It estimates twice the corresponding population Energy-score risk gap, with the stated distance normalization. The squared-mean discrepancy is

$$\widehat{D}_{\mathrm{mean}}(a,b) = \frac{1}{d} \sum_{q=1}^{d} \left[ (\bar{a}_q - \bar{b}_q)^2 - \frac{s_{a,q}^2}{m} - \frac{s_{b,q}^2}{n} \right], \tag{34}$$

where each sample variance has divisor $m-1$ or $n-1$. All negative U estimates are retained. These are two-sample discrepancies, not pointwise prediction errors or thresholded versions of a mean forecast.

Full-path Energy ($E_{\mathrm{full}}$) applies Equation 33 to all $HK = 512$ coordinates. Public-path Energy ($E_{\mathrm{pub}}$) uses the two declared channels only. The three variogram-based statistics apply Equation 34 to fixed $|x_i - x_j|^{1/2}$ features (Scheuerer & Hamill, 2015). Sparse full-channel variogram ($V_{\mathrm{sp}}$) uses all simultaneous unordered channel pairs and each channel's lag-one pairs, including the phase boundary: 4,336 features, not all $(HK)^2$ coordinate pairs. Public variogram ($V_{\mathrm{pub}}$) uses the simultaneous channel pair and all four ordered lag-one pairs within each phase, retaining each time coordinate (152 features). Phase-pooled variogram ($V_{\mathrm{ph}}$) first averages each of those five pair classes within its phase, leaving ten features. The latter is a strong task-matched variogram-based feature score, not a relabeling of the sparse score.

The candidate phase–lag moment statistic ($M_{\mathrm{ph}}$) centers each path's two public channels separately within each phase. Its covariance uses divisor 16; its lag-one matrix uses divisor 15 and no cross-boundary pair. Concatenating $[C_{00}, \sqrt{2}C_{01}, C_{11}, L_{00}, L_{01}, L_{10}, L_{11}]$ over two phases gives 14 features, scored by Equation 34. This per-path centering differs from the earlier ensemble-pooled endpoint. Dynamic-feature Energy ($E_{\mathrm{feat}}$) uses the exact same 14 features in Equation 33; it is a strong transformed-score reference, not an omitted alternative. Finally, pooled covariance ($C_{\mathrm{pool}}$) uses three per-path centered covariance features from the full 32-step block, without phase or lag resolution. It is a blind-spot diagnostic, not a primary comparator. All feature coordinates receive fixed uniform weights. No feature map here identifies every aspect of the complete path law.

### Z.3 Calibration, paired inference and prespecified decision

For every family and method, the threshold is its 1,015th ordered legal calibration statistic, where $1015 = \lceil 1025(1 - 0.01) \rceil$. Rejection is strict exceedance, with no rejection at a tie. Thresholds are stored before challenge generation. This gives a common nominal 1% level, not identical observed false-positive rates or finite-sample conservatism.

The primary average weights eight dynamic conditions equally: half and flat in every family, and scramble in the two AR families. The family weights are therefore $2/8, 3/8, 3/8$, not $1/3$ each. Four legal controls are new in each family and iid scramble. Their 32 method-by-control false-positive intervals use two-sided Clopper–Pearson bounds with each tail allocated $0.025/(2 \cdot 32)$. Together they have at least 97.5% coverage conditional on the fixed thresholds. Candidate qualification requires all four of its upper bounds to be at most 5%.

For each reference method, let $Z_{fi}$ sum the paired candidate-minus-reference rejection indicators over the $k_f = (2,3,3)$ primary conditions in family $f$ and trial $i$. The macro difference is $\widehat{\Delta} = \sum_{f,i} Z_{fi}/(8 \cdot 512)$. Six simultaneous two-sided 97.5% Hoeffding intervals use the predeclared common half-width

$$r = \sqrt{\frac{2(2^2 + 3^2 + 3^2)}{8^2 \cdot 512} \log \frac{2 \cdot 6}{0.025}} = 0.0910494151. \tag{35}$$

Table 65: **All eight evaluator summaries.** Power is the eight-condition average in percent. The false-positive column is the largest of four simultaneous upper bounds, not the observed false-positive rate. "Not established" does not show that a method's true rate exceeds 5%. Abbreviations are defined in the text.

| Statistic | Dynamic power, % | Largest FP upper, % | FP qualification |
|---|---|---|---|
| $E_{\mathrm{full}}$ | 1.099 | 5.218 | Not established |
| $E_{\mathrm{pub}}$ | 0.952 | 4.349 | Established |
| $V_{\mathrm{sp}}$ | 27.051 | 4.933 | Established |
| $V_{\mathrm{pub}}$ | 64.893 | 4.349 | Established |
| $V_{\mathrm{ph}}$ | 79.517 | 4.933 | Established |
| $M_{\mathrm{ph}}$ (candidate) | 80.396 | 3.423 | Established |
| $E_{\mathrm{feat}}$ | 84.595 | 4.349 | Established |
| $C_{\mathrm{pool}}$ (diagnostic) | 0.806 | 4.049 | Established |

Table 66: **All six prespecified paired comparisons.** Differences and simultaneous intervals are in percentage points, candidate minus reference. The practical criterion is a lower bound above 5 pp after candidate false-positive qualification. The last two comparisons remain unresolved.

| Reference | Difference, pp | Simultaneous interval, pp | Practical advantage |
|---|---|---|---|
| $E_{\mathrm{full}}$ | 79.297 | [70.192, 88.402] | Established |
| $E_{\mathrm{pub}}$ | 79.443 | [70.338, 88.548] | Established |
| $V_{\mathrm{sp}}$ | 53.345 | [44.240, 62.450] | Established |
| $V_{\mathrm{pub}}$ | 15.503 | [6.398, 24.608] | Established |
| $V_{\mathrm{ph}}$ | 0.879 | $[-8.226, 9.984]$ | Not established |
| $E_{\mathrm{feat}}$ | $-4.199$ | $[-13.304, 4.906]$ | Not established |

The bound treats each paired trial as a unit; it does not assume that the five variants are independent. The false-positive and difference families jointly have at least 95% coverage by the two $0.025$ allocations. The fixed practical-superiority rule requires candidate qualification and a difference lower bound strictly above five percentage points. Per-condition power intervals, where reported in the complete record, are only pointwise 95% descriptive bounds.

### Z.4 Complete results and complementary blind spots

Tables 65–67 retain every method and condition. The candidate satisfies its false-positive criterion and the practical advantage rule against four references. It does not establish an advantage over phase-pooled variogram or Energy on the same dynamic features. The latter has the highest mean primary power, 84.595%, compared with the candidate's 80.396%. Neither unresolved comparison establishes equivalence or a significant reverse advantage. The macro average also hides hard conditions: candidate power for the half-amplitude change is only 32.422% in Gaussian AR and 15.234% in Student AR. Hence the result supports task-matched measurement, not uniform high power or exclusive superiority of the moment statistic.

Full-path Energy's largest observed legal false-positive rate is 11/512 (2.148%), with simultaneous upper bound 5.218%. It does not establish the prescribed upper bound of 5%, but this is not evidence that its true false-positive rate is above 5%, nor that Energy is an invalid proper score. Its low power against the specified dynamic failures also does not imply population blindness. A common shift gives the reverse comparison: full-path Energy detects 512/512 in every family, while centered moments and difference-based variograms are shift-invariant. Their few rejections match the corresponding legal-control fluctuations. Public-path Energy detects 512/512, 491/512 and 491/512 shifts. The iid scramble remains a legal control, not a detection success.

The result does not license one local statistic to replace full-path quality scoring. The declared phase information helps locate a failure; the shift counterexample demonstrates the cost of invariance. Only two public channels and known phase boundaries are studied. Neither these 14 features nor successful request certificates establish adaptation to unknown change points, all $K = 500$ dependencies, a requested counterfactual law, or real financial, hydrological or weather utility. Finite-

Table 67: **Every method-by-condition count: all 120 cells.** Each entry is a rejection count out of 512. L: legal false-positive control; P: primary dynamic failure; B: predeclared common-shift boundary failure. Counts within a row use the same ensembles, and five rows within a family share trials. They must not be treated as additional independent units.

| Family | Variant | Role | $E_{\text{full}}$ | $E_{\text{pub}}$ | $V_{\text{sp}}$ | $V_{\text{pub}}$ | $V_{\text{ph}}$ | $M_{\text{ph}}$ | $E_{\text{feat}}$ | $C_{\text{pool}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Gaussian iid | New | L | 11 | 6 | 6 | 8 | 9 | 2 | 2 | 7 |
| | Half | P | 9 | 6 | 8 | 42 | 266 | 491 | 501 | 3 |
| | Flat | P | 10 | 6 | 28 | 427 | 512 | 512 | 512 | 5 |
| | Scramble | L | 5 | 8 | 10 | 6 | 10 | 4 | 2 | 7 |
| | Shift | B | 512 | 512 | 6 | 8 | 9 | 2 | 2 | 7 |
| Gaussian AR | New | L | 5 | 6 | 7 | 6 | 9 | 4 | 4 | 5 |
| | Half | P | 5 | 5 | 8 | 114 | 236 | 166 | 248 | 2 |
| | Flat | P | 6 | 3 | 14 | 480 | 511 | 512 | 512 | 3 |
| | Scramble | P | 8 | 11 | 512 | 512 | 512 | 512 | 512 | 5 |
| | Shift | B | 512 | 491 | 7 | 6 | 9 | 4 | 4 | 5 |
| Student AR | New | L | 3 | 4 | 4 | 7 | 8 | 5 | 8 | 6 |
| | Half | P | 3 | 2 | 7 | 104 | 200 | 78 | 156 | 5 |
| | Flat | P | 2 | 2 | 19 | 467 | 508 | 510 | 512 | 4 |
| | Scramble | P | 2 | 4 | 512 | 512 | 512 | 512 | 512 | 6 |
| | Shift | B | 512 | 491 | 4 | 7 | 8 | 5 | 8 | 6 |

bank copying is a separate question and is not used to handicap proper scores in this comparison. No backbone is trained or ranked, and no generation-model SOTA conclusion follows.

The single run used two CPU threads for 31.29 seconds. All 72 raw-input blocks (1.125 GiB) were persisted before their statistics. An independent 66.05-second audit reused every block, checked continuous addresses, seven source texts, finite statistics, threshold ranks, all counts and intervals, and paired contrasts. Maximum statistic and interval discrepancies were $8.71 \times 10^{-15}$ and $4.93 \times 10^{-13}$, respectively, within the unchanged $10^{-10}$ absolute and $10^{-9}$ relative arithmetic tolerances. Saved-input reuse checks include size and modification time, not cryptographic identity. The completed comparison was not retried, expanded or used to retune thresholds, feature maps or corruption amplitudes.

# AA Reproducing the synthetic adoption decision

This specification covers the five-scenario C14 study in Table 4, including its inherited C11 policy fits. It separates the original simulation from arithmetic replay of its saved records. All scored paths have thin dimension $r = 4$, history length $L = 96$, horizon $H = 16$ and $M = 64$ members. The $K = 500$ statement refers to eight registered ambient lift checks from the first replicate; it is not a claim that all scored arrays have 500 independently varying coordinates.

## AA.1 Latent history, base ensemble and independent future

For origin $i$, draw $x_i \sim \text{Unif}([-1,1]^4)$ and a Gaussian-QR orthogonal orientation $O_i$, with QR diagonal signs made positive. Set $S_i = O_i \,\text{diag}(\lambda) O_i^\top$, where $\lambda = (0.65, 0.20, -0.25, -0.60)/\|(0.65, 0.20, -0.25, -0.60)\|_2$. For each of four 24-row history blocks, the covariance is

$$\Sigma_{i\ell} = e^{0.25 x_{i\ell}}(I_4 + 0.45 S_i), \qquad C_{i\ell} = B_{i\ell}(B_{i\ell}^\top B_{i\ell}/24)^{-1/2}\Sigma_{i\ell}^{1/2}, \tag{36}$$

where $B_{i\ell}$ contains iid standard Gaussian entries after column centering. Concatenating the four blocks gives $C_i \in \mathbb{R}^{96\times 4}$ with exact block covariances, up to floating-point error. Matrix roots are symmetric eigendecomposition roots.

Let $P_H = I_H - \mathbf{1}\mathbf{1}^\top/H$. Draw independent Gaussian arrays $Z_{im}, Z_i^* \in \mathbb{R}^{H\times r}$ with entry standard deviation 0.35. Consecutive groups $g$ of eight origins share $a_g = \exp(0.15\eta_g)$, $\eta_g \sim N(0,1)$, independent of those arrays. The base members and one realized future per origin are

$$X_{im}^0 = a_{g(i)} P_H Z_{im}, \qquad Y_i = a_{g(i)} P_H Z_i^* \exp(\beta_{i0} I_4 + \beta_{i1} S_i), \tag{37}$$

where the frozen directional dose is five:

$$\beta_{i0} = 0.25\tanh[(0.08x_{i1} - 0.06x_{i2} + 0.04x_{i3})/0.25]\,,$$
$$\beta_{i1} = 0.90\tanh[5(0.12x_{i1} + 0.36x_{i3} - 0.28x_{i4})/0.90]\,. \tag{38}$$

Thus dose multiplies the raw directional coefficient before saturation. The realized future is not an ensemble member or an exact paired target. True $x_i, S_i, \beta_i$ are not used in fitting or adoption calculations.

### AA.2 Observed shapes and registered distortions

View $v$ observes $C_i^{(v)} = C_i + \sigma_v \rho_i E_i^{(v)}$, where $\rho_i = \sqrt{\|C_i\|_F^2/(96r)}$ and the entries of $E_i^{(v)}$ are iid standard Gaussian. The two noise tapes are independent conditional on the same latent history. Within each observed 24-row block, center the rows to obtain $D_{i\ell}^{(v)}$. The empirical covariance is $V_{i\ell}^{(v)} = D_{i\ell}^{(v)\top} D_{i\ell}^{(v)}/24$. The four policy features are always $\hat{x}_{i\ell}^{(v)} = \log[\mathrm{tr}(V_{i\ell}^{(v)})/r]/0.25$, including in the robust arm.

For the robust shape, cap each centered row radius at its within-block 0.80 quantile: multiply row $d$ by $\min\{1, q_{.80}/\max(\|d\|_2, 10^{-15})\}$, without subsequent recentering. Use NumPy's linear quantile interpolation. Pool the four resulting covariances and normalize their trace-free part:

$$\mathcal{N}(V) = \frac{V - \mathrm{tr}(V)I_r/r}{\|V - \mathrm{tr}(V)I_r/r\|_F}, \qquad \widehat{S}_i^{(v)} = \mathcal{N}\left(\tfrac{1}{4}\sum_{\ell=1}^{4} V_{i\ell,\mathrm{winsor}}^{(v)}\right). \tag{39}$$

The empirical shape uses the unclipped covariances in the same formula. Nonpositive caps or trace scales, or a trace-free norm at most $10^{-12}$, are errors, not alternative estimates.

The five conditions share the same evaluation latent histories, base members and futures. Clean uses noise scales $0.20/0.20$; moderate auxiliary noise replaces only the auxiliary observation with an independent $0.35$ tape; severe primary noise replaces only the primary observation with an independent $0.80$ tape. For reversal, eigendecompose the clean primary's *empirical* normalized trace-free covariance with eigenvalues in ascending order. Orient each eigenvector so its first observed-row projection of magnitude above $10^{-10}$ is positive. If $V_i$ collects these vectors and $J$ is the anti-diagonal identity, set $R_i = V_i J V_i^\top$. Primary reversal uses $C_i^{(1)} R_i$; joint reversal uses $(C_i^{(1)} R_i, C_i^{(2)} R_i)$ with the same primary-derived $R_i$. Recompute all estimates without adding noise. Base members and futures remain unchanged. The common right action preserves the cross-view relation, which is the intended common-cause failure control.

### AA.3 Inherited fit and the two batch decisions

Each of 32 C11 replicates fitted scalar and robust directional policies on 512 origins and activated them using 256 independent validation origins; both splits used primary Gaussian noise 0.20. For $d_i = (1, \hat{x}_{i1}, \ldots, \hat{x}_{i4})$, the fitted map is $\hat{\beta}_{ik} = b_k \tanh[(d_i W)_k/b_k]$, with $(b_0, b_1) = (0.25, 0.90)$. The dose five belongs only to the truth, not this fitted map. The scalar policy has only the first column and fixes $\hat{\beta}_1 = 0$. Its candidate paths, and those of the directional policy, are $X_{im}^0 \exp(\hat{\beta}_{i0} I_4 + \hat{\beta}_{i1} \widehat{S}_i^{(1)})$. No origin-level reliability shrinkage is applied in C14.

Both coefficient matrices start at zero. Each receives 600 float64 CPU AdamW updates, learning rate 0.02, weight decay $10^{-4}$, and gradient norm cap 2. Every update selects 32 origins without replacement, samples 16 target-term members with replacement per origin, and 64 ordered distinct member pairs per origin. Pairs may repeat. The scalar and directional fits use the same sampled tape and minimize the sampled fair Energy in (42). Only terminal checkpoints are admissible. The robust activation bit $A_j$ is one iff validation relative improvement over scalar is at least $0.0005$ and $\bar{d} - t_{.95,31} s_d/\sqrt{32} > 0$, where the 32 values $d$ are consecutive eight-origin means of scalar minus robust Energy. No refit follows. C14 copies these coefficients and bits; all 32 inherited robust bits are one.

Reconstruct the 512 C11 training primary histories and add independent auxiliary histories at noise 0.20. Write $m_\ell, s_\ell$ for training medians and unscaled MADs of primary $\hat{x}_\ell$, and $m_e$ for the median

empirical–robust shape inner product. The old screen for deployment batch $D$ is

$$O_D = \mathbf{1}\left\{ \max_{\ell \le 4} \frac{|\operatorname{med}_D \hat{x}_\ell - m_\ell|}{s_\ell} \le 1.25, \quad m_e - \operatorname{med}_D \langle \widehat{S}_{\mathrm{emp}}, \widehat{S}_{\mathrm{rob}} \rangle_F \le 0.02 \right\}. \tag{40}$$

For each origin define $a_i = \operatorname{clip}_{[-1,1]} \langle \widehat{S}_i^{(1)}, \widehat{S}_i^{(2)} \rangle_F$. Training gives $m_a = \operatorname{med} a_i$ and $s_a = \operatorname{med} |a_i - m_a|$. The cross-view screen and adoption rule are

$$R_D = \mathbf{1}\{(m_a - \operatorname{med}_{i \in D} a_i)/s_a \le 5\}, \qquad A_{j,D}^{\mathrm{old}} = A_j O_D, \qquad A_{j,D}^{\mathrm{combined}} = A_j O_D R_D. \tag{41}$$

Every median in these screens is across the entire 512-origin batch, not a per-origin threshold. All MADs are unscaled; values at most $10^{-8}$ cause an error. Agreement and cross-view drops are signed, without absolute values. The five-MAD threshold was fixed using disjoint development seed 26091401. Adoption uses only the stated past-derived fields. Rejection returns the scalar array already computed from that scenario's primary features, preserving bit-exact fallback within the scenario.

### AA.4 Seeds, score aggregation and scope of uncertainty

For replicate $j = 0, \ldots, 31$, the original C11 seeds are the first uint32 state from NumPy SeedSequence $[26090601, j, k, 0]$: roles $k = 0, 1$ give train and validation truth, $3, 4$ their observation noise, and 5 the optimizer tape. C14 uses $[26091501, j, k]$: roles 0–6 give training auxiliary noise, evaluation truth, clean primary noise, clean auxiliary noise, moderate auxiliary noise, severe primary noise, and gauge replay, respectively. For a truth seed $s$, the base/history generator uses `default_rng(s)`; two children spawned from SeedSequence $[s, 29090241]$ supply the realized-future and eight-origin scale tapes. The original entry points preserve the exact draw order. C14 uses 512 new evaluation origins per policy pair, six CPU workers and one Torch thread per worker.

The full evaluation computes all member pairs, with a complete $H \times r$ path as the score unit:

$$E_i = \frac{1}{\sqrt{Hr}} \left[ \frac{1}{M} \sum_m \|X_{im} - Y_i\|_F - \frac{1}{2M(M-1)} \sum_{m \ne m'} \|X_{im} - X_{im'}\|_F \right]. \tag{42}$$

For scenario $s$, Table 4 pools the 32 saved 512-origin means $\bar{E}_{js}$ as $100[\sum_j \bar{E}_{js}^{\mathrm{policy}} / \sum_j \bar{E}_{js}^{\mathrm{scalar}} - 1]$. This is a ratio of pooled means, not a mean of replicate percentage changes. The old-policy mean selects the saved robust or scalar mean using $A_{j,D}^{\mathrm{old}}$; the combined-policy mean is also saved directly. The auxiliary-moderate and clean policy scores coincide because their primary histories and adopted actions coincide.

The 32 replicates represent repeated fitting and Monte Carlo evaluation under this fixed synthetic law. They are not 32 real datasets; the five paired scenarios do not increase the independent replicate count. As a descriptive check on the saved means, clean and auxiliary-moderate replicate Energy changes have minimum/median/maximum $-0.1782\%/-0.1207\%/-0.0748\%$; joint reversal has $+0.0559\%/+0.0959\%/+0.1611\%$. These are observed spreads, not confidence intervals or independently replicated evidence. Original operating intervals are nominal two-sided 95% Wilson intervals with $z = 1.959963984540054$ ($32/32$ has lower endpoint $0.8928$). Within-replicate effect summaries use eight-origin blocks, not members or coordinates. These summaries neither calibrate the MAD threshold nor establish simultaneous or distribution-free protection. Joint reversal still passes and harms in all 32 replicates.

### AA.5 What the supplied replay reproduces

The archive `paper_a_decision_records_replay_20260926_v1.zip` contains the original records and source, plus a non-training arithmetic replay. From the extracted root, run `python3 -I -S -B scripts/replay_paper_a_decision_records_v1.py`; `--json` also reports the descriptive effect spreads. The original scientific entry points are `run_gauge_equivariant_directional_shape_robustness_v1.py` and `run_gauge_equivariant_directional_relational_observability_v1.py`. Their stochastic, shape-robustness, orientation-witness and relational modules implement the specification above. They have not been rerun for this revision.

The replay reconstructs both screens from saved diagnostics, checks their conjunction with frozen activation, and recomputes Table 4 counts and pooled effects from saved means. This establishes arithmetic reproducibility conditional on those records. It does not regenerate histories, fit policies, resample paths, recompute Energy from individual paths, or independently verify saved pathwise equality and geometry certificates. Full scientific reproduction requires the original generators, fits, observation tapes and numerical dependencies. The saved-record replay is not a new independent confirmation. The saved records do not pin original library versions or implicit AdamW defaults, so numerical or bitwise agreement across versions is not certified. Frozen coefficients retain the specified C14 policies without a new C11 fit.