EXACT FEEDBACK NEED NOT HELP: CORRECTION COMPATIBILITY IN FROZEN NEURAL SIMULATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact observations improve the coordinates they replace yet can worsen an autoregressive neural simulator's later forecasts. We study *correction compatibility*: whether a correction improves a frozen simulator's later forecasts. For exact linear dynamics, harm is possible if and only if the propagated, scored error couples corrected and uncorrected coordinates, so exact partial replacement can hurt even a perfect model (related shock from direct insertion is known in data assimilation), and we show that such harm occurs in released neural simulators. A Rayleigh–Bénard checkpoint that beats persistence by about 40% is harmed by two exact corrections selected as harmful on validation, and the harm replicates, with intervals excluding zero, on a previously unused trajectory from the same test cells; the effect is small (about 1.5% of delayed RMSE) and shown for one checkpoint, an existence result. On Supernova simulations, more exact correction helps on average, and all-channel correction helps in every event; in a post-hoc comparison, a channel-restricted exact correction does worse over one step but better over seven than a fraction-scaled all-channel correction with equal immediate repair. How the corrected state is emitted and fed back is a design choice: development data select opposite interfaces for two acoustic checkpoints, and both hold on 200 untouched trajectories. On a turbulence simulator, with the emission mapping fixed, feeding back a learned, constraint-preserving adjustment improves on discarding it, but the approach harms one other system and adds nothing on another. Feedback should be validated by how it enters recurrence, not by its immediate accuracy.

1 INTRODUCTION

Autoregressive neural PDE models repeatedly consume their own predictions (Brandstetter et al., 2022; Lippe et al., 2023). In a hybrid simulator or observing loop, the next input can instead contain an external correction. Replacing an erroneous coordinate by its exact value necessarily improves that state on the corrected coordinate. It does not follow that a fixed, misspecified network will produce a better future. Learned errors may compensate, and correction changes the input distribution seen by every subsequent recurrent call. We call the corrected state carried back into the history the *feedback*.

The central point is that state accuracy and recurrent compatibility are different properties: a correction can improve the state it modifies while making the simulator's future trajectory worse. We call a correction *compatible* when it improves the closed-loop future. Autonomous competence asks whether a checkpoint beats a baseline; compatibility asks whether a correction improves that checkpoint. On 35 held-out public Rayleigh–Bénard test files, an official checkpoint beats persistence by 40.15% yet is harmed by each of two exact streamfunction corrections selected as harmful on validation, and the harm replicates, with intervals excluding zero, on a previously unused trajectory from the same test cells. We exclude the complete corrected span from the primary error, preventing copied reference values from supplying a trivial improvement. The effect is about 1.5% of delayed RMSE on one checkpoint, so it establishes existence, not prevalence.

Why is delayed value hard to anticipate? A correction reaches the future only through the simulator's response, the difference between corrected and uncorrected rollouts. A counterfactual rollout gives this response, but not whether it cancels or reinforces the unknown future error (Section 2).

Prior data-assimilation work studies observation impact within assimilation systems (Langland & Baker, 2004; Trémolet, 2008; Laloyaux et al., 2025). Here we isolate a different question: whether

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.

an externally supplied correction is compatible with the recurrence of a fixed learned simulator. We freeze a public checkpoint, insert an exact partial state, and select its emitted and carried interfaces on development data before held-out evaluation. The practical rule is to treat both interfaces as part of the simulator and validate them by delayed error over several horizons. Rayleigh–Bénard and Supernova show that immediate repair does not certify delayed value; the acoustic and turbulence studies show that such validation selects different interfaces for different checkpoints.

Our contributions are:

1. **Framing.** We define delayed correction compatibility for frozen neural simulators and state the linear facts that organize our experiments. They are standard in data assimilation, where direct insertion is known to cause imbalance and initialization shock (Hoke & Anthes, 1976; Lynch & Huang, 1992; Ghil & Malanotte-Rizzoli, 1991): exact partial correction can hurt even a perfect linear model exactly when the scored error couples corrected and uncorrected coordinates, and, with the true error moments at one exact correction, a zero-intercept least-MSE update of the uncorrected coordinates cannot hurt on average (the orthogonality principle). Our contribution is not these linear facts but evidence that exact-correction harm, and its dependence on how a correction is emitted and fed back, occurs in released neural simulators under prespecified protocols.
2. **Controlled evidence.** On Rayleigh–Bénard, a released checkpoint that beats persistence is harmed by two exact corrections, and the harm replicates, with intervals excluding zero, on a previously unused trajectory from the same test cells. On Supernova simulations, a sweep of exact channel corrections shows that more correction helps on average but not event by event; a post-hoc matched-repair test, whose arms differ in direction and information content, shows that equal immediate repair need not give equal delayed value.
3. **Interface validation.** Development data select opposite recurrent interfaces for two released acoustic checkpoints, and both choices hold on untouched trajectories from the same simulator. On a turbulence simulator, carrying back a protected learned adjustment improves on discarding it, but the approach harms Shear Flow and adds nothing on Viscoelastic.

Section 2 develops the theory and Section 3 the interventions; Section 4 tests them on Rayleigh–Bénard (4.1), Supernova (4.2 and 4.3), and acoustic and turbulence simulators (4.4 and 4.5), and reports failed remedies (4.6).

2 Why exact correction can hurt

2.1 Measuring delayed correction value

The results in this section are elementary and known in data assimilation; we state them in the form our experiments use. We compare two rollouts of the same frozen simulator, an intervention arm a and a declared reference arm b, and ask whether the intervention has lower delayed error on a fixed set of scored coordinates. Let F be a frozen autoregressive simulator, x_t the reference state, and y_t^a, y_t^b the arms' scored pre-analysis forecasts. Let P be the orthogonal projector onto the coordinates changed by a fixed correction. For a family $\{P_A\}$, let P_{all} project onto the union of all predeclared corrected spans. The scoring projector Q is predeclared and fixed across the arms; it may exclude corrected coordinates (scoring only the uncorrected, or *complement*, coordinates) and fixed gauge or nuisance coordinates. We use, for example, $Q = I - P$ for a fixed correction, $Q = I - P_{\text{all}}$ when excluding an entire candidate family, and $Q = I$ for a common full-state endpoint. For insertion fraction $g \in [0, 1]$ ($g = 1$ is exact insertion, $g < 1$ damped insertion), the analyzed state (the *analysis*) is

$$z_t^g = y_t^g + gP(x_t - y_t^g), \tag{1}$$

and under *direct carry* z_t^g rather than y_t^g enters the next history (with $g = 1$, direct insertion in data-assimilation terms; repeated with $g < 1$, a discrete form of nudging (Hoke & Anthes, 1976)); Section 3 defines interfaces that alter emission or carry. In every study the forecast is scored before that step's correction. For trajectories in a declared aggregation unit $\mathcal{G}$ and future window T, define

$$\mathcal{V}_Q(a; b, T \mid \mathcal{G}) = 100 \left(1 - \sqrt{\frac{\sum_{n \in \mathcal{G}} \sum_{t \in T} \|Q(y_{n,t}^a - x_{n,t})\|_2^2}{\sum_{n \in \mathcal{G}} \sum_{t \in T} \|Q(y_{n,t}^b - x_{n,t})\|_2^2}} \right). \tag{2}$$

Positive delayed value means reduced normalized RMSE. Rayleigh uses autonomy as b and all files as one pooled unit; TGC uses the direct analyzed-history arm as b, computes the value within each fold-seed role, and averages the six dependent roles equally. Autonomous competence is defined analogously against persistence using the full state. Competence compares two forecasts; compatibility compares two input distributions presented to the same recurrent model.

For actions that correct different full channels, using $Q = I - P_A$ would change the scoring coordinates with the action. The Supernova correction sweep therefore uses a common full-state delayed endpoint. Error under a fixed Q that excludes the corrected span, which we call *complement error*, suits a fixed correction and is our primary counterexample metric.

2.2 When partial correction is safe

For any paired baseline and corrected branches, stack the scored future window and write

$$e = Q(y^{\text{base}} - x), \qquad r = Q(y^{\text{corr}} - y^{\text{base}}).$$

Then, without linearizing F,

$$\underbrace{\|e + r\|_2^2 - \|e\|_2^2}_{\Delta\mathcal{E}_Q} = \underbrace{2\langle e, r\rangle}_{\text{signed alignment}} + \underbrace{\|r\|_2^2}_{\text{response energy}} . \tag{3}$$

Here $\Delta\mathcal{E}_Q > 0$ is harm, the opposite sign convention from the positive-is-good values reported below. The response-energy term always adds error, so a correction helps only if $-2\langle e, r\rangle > \|r\|^2$: a response can point partly toward the reference and still be harmful. The response alone cannot certify the sign: the same r gives $\Delta\mathcal{E}_Q = \|r\|^2$ if $e = 0$ and $-\|r\|^2$ if $e = -r$, so a predictor must also use data relating response geometry to future error.

Harm does not require nonlinear learned dynamics. For exact linear dynamics, it is possible exactly when the scored future couples errors inside and outside the corrected span. Let $W \succeq 0$ be the finite-horizon error metric induced by the propagator and scoring rule. Write the current error as e_0 and decompose it as $p + q$, with $p = Pe_0$ and $q = (I - P)e_0$. Removing p improves squared error by

$$p^\top W p + 2p^\top W q. \tag{4}$$

Proposition (universal linear compatibility). Assume P is an orthogonal projector and W is symmetric positive semidefinite. Exact removal of the P component is non-harmful for every current error if and only if

$$PW(I - P) = 0. \tag{5}$$

The proof is in Appendix G. The proposition assumes a perfect linear model; with model bias, a term linear in p can cause harm even when $PW(I - P) = 0$. This characterization explains when harm is possible, but it is not a target-free rule for predicting harmful corrections in a nonlinear simulator.

Universal compatibility is deliberately stringent: it asks that the correction never hurt, whatever the current error. A weaker and often more relevant question is whether it helps on average over a physical population. For fixed P and W, define $I(e_0) = e_0^\top W e_0 - ((I - P)e_0)^\top W((I - P)e_0)$. For a population of current errors with finite second moment $S = \mathbb{E}[e_0 e_0^\top]$, the same calculation gives the exact average-case identity

$$\mathbb{E}[I(e_0)] = \operatorname{tr}(PWPS) + 2\operatorname{tr}(PW(I - P)S). \tag{6}$$

The first term is the propagated energy removed from the corrected span. The second is signed cross-span alignment and can overturn it; it depends on the error cross-moment (the cross-covariance when the relevant errors are mean-zero). Equation (6) is a characterization under the population encoded by S, not a target-free certificate: estimating its signed term requires reference-dependent current-error statistics. The identity motivates the signed probe in Section 4.1. If the uncorrected coordinates are also updated with the zero-intercept least-MSE linear gain $K = S_{qp}S_{pp}^{+}$ from the uncentered second moments, $q \mapsto q - Kp$, the expected improvement becomes $\operatorname{tr}(W[I; K]S_{pp}[I; K]^\top) \geq 0$ for every $W \succeq 0$, where $[I; K]$ maps the corrected span into the full state. Its improvement is at least that of exact insertion; the difference is $\operatorname{tr}(W_{qq}KS_{pp}K^\top)$ with $W_{qq} = (I - P)W(I - P)$ (Appendix G). Average harm under a perfect linear model therefore requires a cross-moment that the correction ignores; the bound needs one exact correction, a perfect linear model, and the moments at

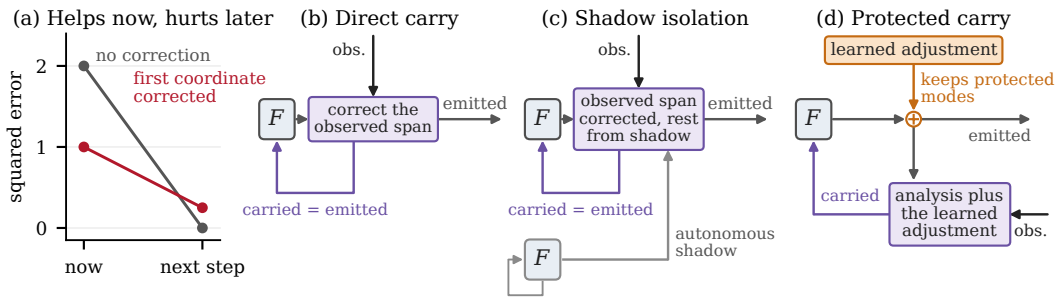

Figure 1: Why correction can hurt, and three recurrent interfaces. (a) The exact linear example of Section 2: correcting the first coordinate halves the current squared error but raises the next-step error from 0 to 1/4. (b) Direct carry scores the forecast before the event correction and returns the resulting analysis directly to the history. (c) Shadow isolation runs a second copy of the frozen F as an autonomous shadow; outside the observed span, both the emitted and the carried state come from the shadow. (d) Protected carry adds a learned adjustment that leaves the protected modes untouched and carries back the analysis plus this adjustment; the matched emit-only arm carries back the analysis of the forecast alone.

that correction time, and holds only on average. The Rayleigh baseline of Section 4.1 fails the key conditions (its gain is cycled, pooled over leads, and applied to a biased nonlinear host), so the bound does not attribute the Rayleigh harm to missing cross-moment structure.

A two-dimensional example makes the cancellation concrete (Figure 1a). Let $M = \left(\begin{smallmatrix} 0 & 0 \\ 1/2 & 1/2 \end{smallmatrix}\right)$ be exact one-step dynamics, $e_0 = (1, -1)^\top$, and let P correct the first coordinate. The correction reduces current squared error from 2 to 1. Yet $Me_0 = (0, 0)^\top$ has no future error, whereas $M(I - P)e_0 = (0, -1/2)^\top$ has error $1/4$, under full-state and complement scoring alike. The exact partial update removes favorable cross-coordinate cancellation even with a perfect linear model.

Insertion fraction and horizon. For a fraction g of the exact update the improvement is $I(g) = 2g\, p^\top W(p+q) - g^2 p^\top W p$, so a small step helps whenever $p^\top W(p+q) > 0$ even if the exact update hurts; we therefore evaluate prespecified fractions (Appendix G). Repeated after every prediction, damped insertion acts as relaxation, which this one-step identity does not describe. For a linear propagator M and scored horizons $1{:}L$, the matrix in the proposition is

$$W_L = \sum_{\ell=1}^{L} (M^\ell)^\top Q^\top Q M^\ell, \tag{7}$$

a finite-horizon observability Gramian for the scored error. Dynamics can make $PW_L(I - P)$ nonzero by transporting error across channels, so compatibility can change with horizon without any change in the observation. Replacing M^ℓ by a rollout Jacobian gives a local diagnostic for a nonlinear simulator, not a certificate, so we report every prespecified horizon. All-channel exact insertion is automatically compatible when the reference follows the model's exact dynamics; for a misspecified learned simulator it is an empirical repair test (Section 4.2).

3 Interventions and Evaluation

At an event, distinguish the pre-analysis forecast scored before correction, the post-event exposed state, and the state fed into the simulator's next history. We call the mapping that fixes the latter two the *recurrent interface* (Figure 1). A *correction* is the update applied at an *event*, a time at which observations arrive, set by its support and insertion fraction; an *intervention arm* fixes the correction together with emission, carry, timing, and scoring; and the *protocol* is the complete prespecified experimental specification. Each system declares its own corrected or protected span (P or P_A) and scoring projector Q; scores are therefore not comparable across systems.

Direct carry. The pre-analysis forecast is scored before the event correction, and the resulting analysis enters the recurrent history directly, without isolation or a learned carry transformation

(Figure 1b). Rayleigh–Bénard applies it after every prediction and Supernova once per event; it is also the classic U-Net's Acoustic interface and the TGC analyzed-history baseline.

Shadow isolation. The acoustic construction uses no learned adapter. Let P project pressure onto 16 prespecified, geometry-conditioned modes (fixed within a trajectory) and $Q = I - P$, including both velocity fields in Q. An autonomous shadow predicts $b_{t+1} = F(b_{t-3:t})$. A corrected branch first predicts $d_{t+1} = F(c_{t-3:t})$, then emits and feeds back

$$c_{t+1} = Pd_{t+1} + Qb_{t+1}, \qquad c_{t+1} \leftarrow Px_{t+1} + Qb_{t+1} \quad \text{at events.} \tag{8}$$

Its complete Q component equals autonomy by construction at every scored state, whereas direct carry appends the event analysis. Isolation requires two concurrent rollouts and exact current event coefficients (basis construction in Appendix A.2).

Protected carry. On turbulence with gravitational cooling (TGC), both arms use the same emission mapping and differ only in whether a learned adjustment stays in the carried history. At events, an analysis A_{x_t} sets three observed channel means of a predeclared 24-mode basis to their values; the adjustment R, the modified-minus-direct output of a compact trained input stem on the frozen host, may not change these 24 protected coefficients. With forecast $d_t^j = F(h_t^j)$, constrained adjustment r_t^j, emitted state $o_t^j = d_t^j + r_t^j$, and analyses $a_t^j = A_{x_t}(d_t^j)$ and $\tilde{a}_t^j = A_{x_t}(o_t^j)$, the carried states of the emit-only and protected-carry arms are

$$s_t^{\text{emit}} = a_t^{\text{emit}}, \qquad s_t^{\text{carry}} = a_t^{\text{carry}} + \Pi_A(\tilde{a}_t^{\text{carry}} - a_t^{\text{carry}})\,, \tag{9}$$

where Π_A keeps the adjustment outside the protected span, so the carried state keeps the observed coefficients. Their paired contrast estimates the total downstream value of carry, including mediated effects; a norm-matched but spatially misaligned (parity-shuffled) carry is a descriptive control. The projections, constraints, architecture, and training are in Appendix B.

3.1 Evaluation protocol

All branches start from the same pre-event history, use the same frozen host, and are scored on common delayed coordinates and time windows. Let $E_{i\ell}^{A,g}$ be squared error on the comparison's declared scoring coordinates at delay ℓ. For event i and horizon L, per-event utility is

$$U_{i,A,g,L} = 100\left(1 - \sqrt{\frac{\sum_{\ell=1}^{L} E_{i\ell}^{A,g}}{\sum_{\ell=1}^{L} E_{i\ell}^{\text{auto}}}}\right). \tag{10}$$

Negative utility is harm. Acoustic instead prespecified an *SSE* (sum of squared errors) benefit, a ratio of sums, $B_{m:c} = 100(\sum_{i,t} E_{it}^{c} - \sum_{i,t} E_{it}^{m})/\sum_{i,t} E_{it}^{\text{auto}}$ in points (percent of autonomous SSE) over complete trajectories and 13 delayed steps; it is not numerically comparable with $\mathcal{V}_Q$ or U. All three are positive-is-better percentages of a reference arm's error ($\mathcal{V}_Q$ and U of RMSE, B of SSE), and contrasts between arms are reported in points. Uncertainty units are complete files or trajectories, as stated per study.

A result is *prospective* when its protocol, code, data identities, and aggregation were frozen before the evaluated files or trajectories were read, *post hoc* when it analyzes units already read (even with an analysis specified before it ran), and *exploratory* for unadjusted variants and development screens. Appendix A.1 gives the formal estimands, resampling schemes, and dependence sensitivity, and Table 1 (Appendix A) lists every headline study with its status.

4 Experiments

We test four consequences of correction compatibility: whether exact correction can hurt, how support and horizon affect delayed value, whether the preferred interface depends on the checkpoint, and whether carry matters with emission fixed. All data and checkpoints are public assets from The Well (Ohana et al., 2024); Appendix A gives full protocols.

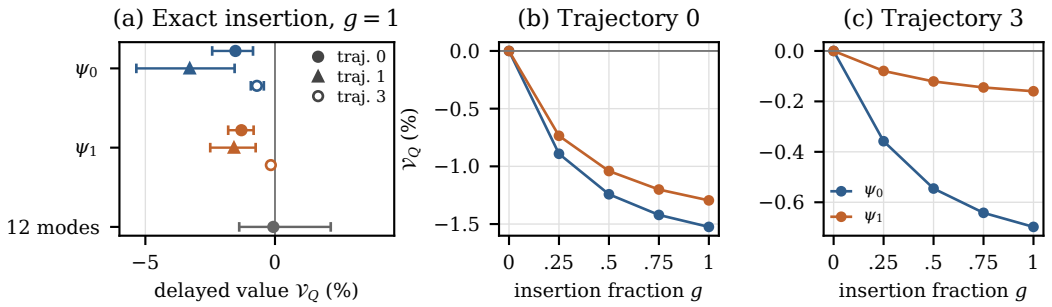

Figure 2: A competent Rayleigh–Bénard checkpoint (it beats persistence by 40.15% $[27.13, 52.62\%]$) is harmed by exact correction. (a) Exact insertion of either streamfunction-mean mode lowers the delayed value on trajectory 0 (prospective test), trajectory 1 (test-informed follow-up), and trajectory 3 (prospective replication), whereas correcting the complete 12-mode span shows no detectable harm (trajectory 0). Intervals resample 35 complete test files. (b, c) Damped insertion ($g < 1$; point estimates, with $g = 0$ autonomy by definition) reduces but does not remove the harm at every tested $g \geq 0.25$.

4.1 A COMPETENT SIMULATOR CAN BE HARMED BY EXACT CORRECTION

Design. In this prospective test, the official ConvNeXt U-Net is evaluated on trajectory 0, the first native trajectory, of each of 35 held-out Rayleigh–Bénard test files. Two corrections of the streamfunction-mean modes ψ_0 and ψ_1 (the two horizontally uniform velocity modes of a 12-mode physical basis; Appendix A.2), selected as harmful in development on 15 validation clusters, were fixed before test access. Each is inserted exactly after every prediction with direct carry (exact insertion), and raw predictions are scored on steps 6 to 12 outside the pressure mean (a gauge freedom) and the complete 12-mode span, with files as uncertainty units. The protocol's primary claim, that a one-step probe ranks the 12 modes with Spearman $\rho \geq 0.70$, failed; harm replication was a prespecified secondary condition.

On this trajectory-0 cohort, the checkpoint improves over persistence by 40.15% ($[27.13, 52.62\%]$), yet both selected modes have negative delayed value $\mathcal{V}_Q$ when pooled over the 35 test files: -1.52% ($[-2.42, -0.85\%]$) and -1.29% ($[-1.80, -0.82\%]$) (Figure 2a). Each is harmful on 21/35 files (two-sided sign test $p = 0.31$), so the harm is a pooled effect whose file-level majority is not significant. Their full-state values are also negative, and correcting the complete 12-mode span gives an interval crossing zero. Both modes were fixed before test access, so no test-time selection is involved; none of the other ten modes has an interval wholly below zero. A signed probe from Equation (6), chosen on validation data, flags both harmful modes on test but misses its prespecified ranking target, so it is not a selector (Appendix F). The claim is existence, not universal harm. A later test-informed follow-up on trajectory 1 found the same harms (-3.29% and -1.58%; Appendix C). In a final test-informed, prospectively frozen check on outcome-unseen trajectory 4, both modes again have negative complement utility, but autonomy is worse than persistence and thus fails the required competence gate; we do not count it as confirmation (Appendix C).

Replication and a withdrawn baseline. On trajectory 3 of each test file, not used before, the harm of exact insertion replicates with smaller magnitude, and both intervals exclude zero (-0.70% $[-0.93, -0.43\%]$ and -0.16% $[-0.22, -0.09\%]$). Damped insertion ($g < 1$) reduces but does not remove it (Figure 2b,c). A post-hoc lead-pooled uncentered regression gain that spreads each correction into the uncorrected state failed its prespecified skill check (cycled on validation, it multiplied the post-analysis complement error by 32 and 4.8), so it does not test covariance-aware correction and we do not interpret its results (Appendix C).

4.2 MORE EXACT CORRECTION HELPS ON AVERAGE, NOT EVENT BY EVENT

Design. The prospective Supernova study evaluates the official four-frame 64^3 U-Net on 16 untouched files from two mass regimes (4 and 12 files), giving 96 events. At each event, an exact correction of channel subset A with fraction g replaces the newest predicted frame by

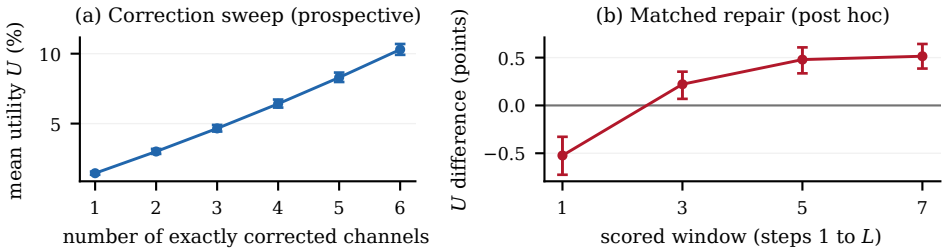

Figure 3: Supernova sweep and matched repair. (a) Seven-step mean utility over exact subsets of each cardinality. (b) Channel-restricted exact update minus a fraction-scaled all-channel update matched for immediate SSE repair; positive values favor the channel-restricted update, which loses at $L = 1$ but wins from $L = 3$ onward. Intervals resample 16 complete files within mass (96 events).

$z_{A,g} = \hat{x}_0 + gP_A(x_0 - \hat{x}_0)$. The sweep evaluates all 63 nonempty subsets at $g = 1$, plus $g \in \{0.25, 0.5, 0.75\}$ for each singleton, the thermodynamic and kinematic three-channel blocks, and all six channels: 90 arms per event, none selected, all sharing the same history; every arm preserves unselected channels bitwise and strictly reduces event-time full-state SSE. Support and correction amount vary together. The primary $L = 7$ comparison is all-channel exact insertion against the event-wise hindsight best of the 62 proper exact subsets, with complete files as uncertainty units.

At $L = 7$, mean utility over all exact subsets of each size and both masses rises steadily with support, from 1.45% for single-channel updates to 10.30% for all six channels (Figure 3a), while the share of harmful arms falls from 11.34% to zero. Larger support helps on average but not event-wise: prespecified amount and nested-support controls show that more full-frame correction helps, while adding an exact channel is not uniformly beneficial, and a post-hoc proportional benchmark attributes most of the trend to the immediate SSE removed (Appendix A.3). The hindsight best subset uses future reference states and is not deployable; at $L = 7$, all-channel exact insertion still exceeds it by 0.27 points $[0.10, 0.43]$, with the largest gap at $L = 3$. This is a population-mean contrast, not pointwise dominance: a proper subset wins 29 of 96 events and four of 16 file means are negative (paired file-level sign-flip $p = 0.035$, post hoc). All-channel insertion helps in all 96 events, including those with harmful proper subsets, but this is an observed repair, not a deployable method or proof that incomplete support caused the failures. The gap has the same sign in both mass regimes.

4.3 Equal immediate repair does not imply equal delayed value

This post-hoc follow-up uses the same 16-file Supernova corpus. The matched-repair test pairs each proper exact subset with a fraction-scaled all-channel correction that removes the same immediate SSE. Averaged over the 62 proper subsets, exact-subset minus matched all-channel utility reverses from -0.52 points $[-0.72, -0.33]$ at $L = 1$ to 0.51 $[0.39, 0.64]$ at $L = 7$, becoming positive by $L = 3$ (Figure 3b), and the reversal persists under every leave-one-file-out analysis. Equal scalar repair therefore does not make two recurrent inputs dynamically equivalent. The arms differ in direction, information content, norm, and exactness pattern, so the test does not identify an effect of support alone; Appendix Table 3 gives cardinality contrasts with the multiplicity-adjusted tests prespecified in its protocol. A proportional benchmark and a post-hoc six-file increment-spreading comparator appear in Appendix A.3.

4.4 The preferred recurrent interface depends on the checkpoint

Design. The Acoustic study uses the official classic and ConvNeXt U-Nets. One 100-trajectory development file selects direct carry for the former and shadow isolation for the latter; an untouched 100-trajectory qualification file must confirm both choices before a prospective 200-trajectory test on two further files. Repeated joint 16-mode pressure observations are scored on 13 delayed pre-analysis states against autonomy and the unselected interface.

Both selected interfaces beat autonomy, the unselected interface, and an events-only complement reset, with the same direction in each test file and under maze-cluster resampling (Figure 4; Appendix Table 2). Their SSE benefit over autonomy is 7.33 points $[6.82, 7.86]$ (classic U-Net) and 1.04

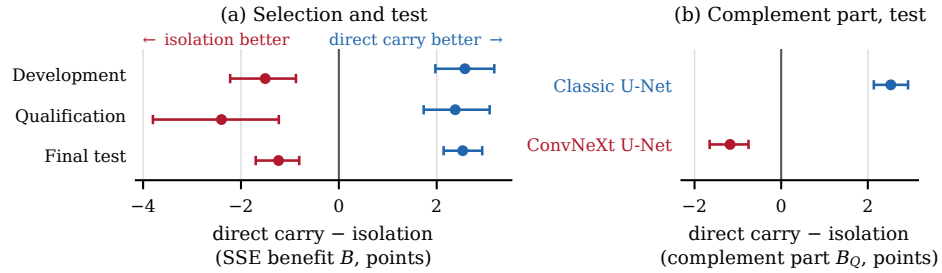

Figure 4: The preferred recurrent interface depends on the checkpoint (blue: classic U-Net; red: ConvNeXt U-Net). (a) Direct carry minus shadow isolation SSE benefit across three disjoint phases. (b) On final test, direct carry lowers complement error for classic U-Net but raises it for ConvNeXt; shadow isolation has zero complement part B_Q by construction. Benefits are points of autonomous full-state SSE; intervals are complete-trajectory bootstraps, stratified by file on final test. The stable crossover supports checkpoint-level selection, not a universal interface.

$[0.90, 1.20]$ (ConvNeXt U-Net). The difference is located in the complement. Shadow isolation makes the complement error equal the autonomous shadow's by construction, whereas direct carry changes the complement part by 2.53 points $[2.14, 2.93]$ for classic U-Net but -1.18 $[-1.65, -0.75]$ for ConvNeXt. Isolation therefore blocks a harmful complement effect for ConvNeXt and removes a beneficial one for classic U-Net. A post-hoc emit-only blend, which emits the shadow's complement but carries the corrected state back, scores 0.99 points for ConvNeXt against 1.04 for isolation: recurrence creates the complement harm, but emitting the shadow's complement removes most of it, and isolating the carry adds only 0.05 points $[0.03, 0.08]$. Section 4.5 varies carry alone. A post-hoc response decomposition, and a local Jacobian check that failed its validity checks, are in Appendix D.

4.5 Protected carry helps with the emission mapping fixed

Design. Disjoint public development files fit the compact TGC adapter and protection operators. The prospective test uses six untouched files and six modules from two folds and three seeds. Its primary matched contrast holds the emission mapping fixed and compares the emit-only adjustment with protected carry.

The direct-carry baseline, which carries the analysis of the unmodified forecast, beats persistence. Relative to it, protected carry improves the complement value by 1.51% $[1.07, 2.10]$ and the full-state value by 1.14% $[0.81, 1.62]$. Against the emit-only adjustment, protected carry adds 0.97 points $[0.68, 1.36]$ of complement value, and this complement contrast is positive for all six dependent procedure replicates. Protected carry also has a better full-state point estimate than emit-only (descriptive) (Figure 5). The parity-shuffled carry does worse than the emit-only arm, so the adjustment's norm alone does not explain the improvement, and every prespecified check passes. No autonomous (no-analysis) arm was run, so the value of analysis itself on TGC is unmeasured.

4.6 Failed remedies and transfer limits

Three further routes failed (Appendix D); the selector audit is post hoc and the others are exploratory screens. A selector that chooses channels from the counterfactual response of each correction did no better than simpler scorers. Supernova affine completion, a development-only regression-gain screen that spreads the revealed residual into unobserved channels at each voxel, recovered only 9.9% of the full-state opportunity and was never run on test. The TGC approach transferred negatively to Shear Flow (protected-carry value -24.4% over six files) and added no pooled carry value on Viscoelastic.

5 Related work

Data-assimilation shock and observation impact. Direct insertion can unbalance a forecast; nudging, digital-filter initialization, and incremental analysis update reduce such shock (Hoke & Anthes, 1976; Lynch & Huang, 1992; Bloom et al., 1996), while observation impact can assign

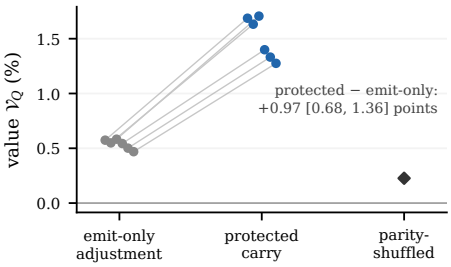

Figure 5: Prospective TGC carry test: each line is one learned module (two folds, three seeds), in complement value $\mathcal{V}_Q$ relative to the direct-carry reference. The annotation is the prespecified interval for $\Delta\mathcal{V}_Q$; the pooled parity-shuffled carry is a descriptive control.

signed downstream value (Gelaro et al., 2010). Kalman-style analyses spread increments through error cross-covariances (Lorenc, 1986; Evensen, 2003), and low-mode insertion can asymptotically synchronize a perfect dissipative model (Hayden et al., 2011; Azouani et al., 2013). We instead test whether an external correction is compatible with a fixed learned recurrence. Our post-hoc Rayleigh regression gain fails its skill check and the conditions of Section 2, so flow-dependent, cycle-consistent, and bias-aware gains and operational IAU remain untested.

Neural simulator correction. Long rollouts expose distribution shift and accumulation (Brandstetter et al., 2022; Lippe et al., 2023; Koehler et al., 2024). Prior work learns forecast-analysis recursion or correction inside hybrid solvers (Singh et al., 2024; Wei et al., 2025), triggers external solvers (Srikishan et al., 2025; Roy et al., 2025), or routes refinement for frozen PDE models (Li et al., 2026). Forecast-analysis cycling, projection, and separate adapter locations are not individually new. Our setting is exposure bias (Bengio et al., 2015; Ranzato et al., 2015) in reverse: inserting exact observations changes the next-input distribution.

Decision context. Selective prediction, deferral, and OOD scores flag unreliable predictions (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019; Madras et al., 2018; Hendrycks & Gimpel, 2016); compatibility instead assigns delayed value to supplied corrections, as observation impact does (Langland & Baker, 2004; Trémolet, 2008; Laloyaux et al., 2025).

6 Discussion, Limitations, and Conclusion

Discussion. Whether a correction helps depends on the checkpoint and on how the corrected state re-enters the history, not only on how much error it removes now. When validating, vary correction amount as well as corrected coordinates, and give emission and carry ablations the same timing, scoring coordinates, and windows. The studies separate these issues in sequence: Rayleigh establishes that harm can occur despite autonomous skill, Supernova separates immediate repair from delayed value, and Acoustic and TGC isolate the recurrent interface. Together, they make the checkpoint, correction, and interface the relevant unit of validation rather than the correction alone.

Limitations. Results are host- and correction-specific, and interfaces are selected offline. Supernova has 16 files and one checkpoint, and its matched-repair test is post hoc. Rayleigh demonstrates existence on one checkpoint; its 35 test files support four sequential studies, with initial-state overlap across validation and test cells. TGC has six test files and transfers negatively; Acoustic has one system and two checkpoints. Noisy, delayed, or costly observations remain open, and no intervention certifies stability or conservation.

Conclusion. State accuracy and recurrent compatibility differ. Exact correction can hurt, and delayed value depends on support, horizon, checkpoint, emission, and carry. These interfaces require held-out rollout validation. Immediate repair alone is therefore not a sufficient selection criterion. A practical evaluation should select interfaces on development rollouts and report their delayed utility on untouched trajectories.

AI USE STATEMENT

In this work, we used generative AI tools for literature discovery, development of the conceptual framework and hypotheses, design of methods and experiments, implementation and testing of methods and analysis code, execution and analysis of experiments, interpretation of results, preparation of scientific figures, formulation and checking of mathematical claims and proofs, and drafting and editing the manuscript. We have not used generative AI tools to generate research datasets, translate text, clean or reformat datasets, or perform qualitative or thematic data analysis; those required disclosure categories are not otherwise applicable to this work. We have reviewed all AI-assisted work: the authors verified the reported code, calculations, citations, claims, and artifacts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly released simulation datasets and model checkpoints (The Well; Ohana et al., 2024). It involves no human subjects, personal data, or crowdsourcing. We see no direct potential for harm; the results argue for validating corrected neural simulators before relying on them.

REPRODUCIBILITY STATEMENT

Appendices A and C specify the data partitions, uncertainty units, intervention inventories, and prospective versus post-hoc status of the headline studies, including every use of the Rayleigh test files. The supplementary artifact contains immutable sufficient statistics, portable aggregation replays, figure generators, tests, and exact identities for the released datasets and checkpoints. It regenerates figures and generated TeX from authenticated summaries, reaggregates designated summaries, and hash-authenticates the remaining canonical summaries; the reproduction guide identifies each category. The code and evidence are available anonymously at `https://anonymous.4open.science/r/correction-compatibility-7AF1/README.md` and will be released publicly upon publication.

REFERENCES

Abderrahim Azouani, Eric J. Olson, and Edriss S. Titi. Continuous data assimilation using general interpolant observables. *Journal of Nonlinear Science*, 24:277 – 304, 2013. URL https://api.semanticscholar.org/CorpusID:6126553.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *ArXiv*, abs/1506.03099, 2015. URL https://api.semanticscholar.org/CorpusID:1820089.

Stephen C. Bloom, Lawrence Takacs, Arlindo M. da Silva, and David V. Ledvina. Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124:1256–1271, 1996. URL https://api.semanticscholar.org/CorpusID:124307650.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural pde solvers. *ArXiv*, abs/2202.03376, 2022. URL https://api.semanticscholar.org/CorpusID:246634432.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010. URL https://api.semanticscholar.org/CorpusID:10773394.

Geir Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003. URL https://api.semanticscholar.org/CorpusID:129233333.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:59316904.

Ronald Gelaro, Rolf H. Langland, Simon Pellerin, and Ricardo Todling. The thorpex observation impact intercomparison experiment. *Monthly Weather Review*, 138:4009–4025, 2010. URL https://api.semanticscholar.org/CorpusID:123084990.

Michael Ghil and Paola Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. *Advances in Geophysics*, 33:141–266, 1991. URL https://api.semanticscholar.org/CorpusID:118326803.

Kevin Hayden, Eric Olson, and Edriss S. Titi. Discrete data assimilation in the lorenz and 2d navier–stokes equations. *Physica D: Nonlinear Phenomena*, 240(18):1416–1425, 2011. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2011.04.021. URL https://www.sciencedirect.com/science/article/pii/S016727891100114X.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2016. URL https://api.semanticscholar.org/CorpusID:13046179.

James E. Hoke and Richard A. Anthes. The initialization of numerical models by a dynamic-initialization technique. *Monthly Weather Review*, 104:1551–1556, 1976. URL https://api.semanticscholar.org/CorpusID:124570425.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thürey. Apebench: A benchmark for autoregressive neural emulators of pdes. *ArXiv*, abs/2411.00180, 2024. URL https://api.semanticscholar.org/CorpusID:273798282.

Patrick Laloyaux, Mihai Alexe, Eulalie Boucher, Peter Lean, Ewan Pinnington, Simon Lang, Tobias Necker, and Anthony McNally. Using data assimilation tools to dissect graphdop. *arXiv preprint arXiv:2510.27388*, 2025. URL https://api.semanticscholar.org/CorpusID:282719068.

Rolf H. Langland and Nancy L. Baker. Estimation of observation impact using the nrl atmospheric variational data assimilation adjoint system. *Tellus A: Dynamic Meteorology and Oceanography*, 56:189 – 201, 2004. URL https://api.semanticscholar.org/CorpusID:121519656.

Chengze Li, Lingwei Wei, Li Sun, Hongbo Lv, Jie Yang, Hanrong Zhang, Kening Zheng, Wei-Chieh Huang, Enze Ma, and Philip S. Yu. ARC-STAR: Auditable post-hoc correction for PDE foundation models. *arXiv preprint arXiv:2605.22222*, 2026. URL https://api.semanticscholar.org/CorpusID:288650466.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. Pde-refiner: Achieving accurate long rollouts with neural pde solvers. *ArXiv*, abs/2308.05732, 2023. URL https://api.semanticscholar.org/CorpusID:260775609.

Andrew C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112:1177–1194, 1986. URL https://api.semanticscholar.org/CorpusID:119759395.

Peter Lynch and Xiang-Yu Huang. Initialization of the hirlam model using a digital filter. *Monthly Weather Review*, 120:1019–1034, 1992. URL https://api.semanticscholar.org/CorpusID:124470884.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf.

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer,

and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=00Sx577BT3.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732, 2015. URL https://api.semanticscholar.org/CorpusID:7147309.

Rajyasri Roy, Dibyajyoti Nayak, and Somdatta Goswami. Anchor: Error-controlled adaptive numerical correction for neural operator time marching. *arXiv preprint arXiv:2512.19643*, 2025. URL https://api.semanticscholar.org/CorpusID:289299688.

Ashutosh Singh, Ricardo Augusto Borsoi, Deniz Erdoğmuş, and Tales Imbiriba. Learning semi-linear neural operators : A unified recursive framework for prediction and data assimilation. *ArXiv*, abs/2402.15656, 2024. URL https://api.semanticscholar.org/CorpusID:267938654.

Bharat Srikishan, Daniel O'Malley, Mohamed Mehana, Nicholas Lubbers, and Nikhil Muralidhar. Model-agnostic knowledge guided correction for improved neural surrogate rollout, 2025. URL https://arxiv.org/abs/2503.10048.

Yannick Trémolet. Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 60:964 – 978, 2008. URL https://api.semanticscholar.org/CorpusID:119773067.

Hao Wei, Aleksandra Franz, Björn Malte List, and Nils Thuerey. INC: An indirect neural corrector for auto-regressive hybrid PDE solvers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=s3Uk3lrfjy.

A ESTIMANDS, PROTOCOLS, AND SECONDARY RESULTS

Table 1: Studies at a glance. $\mathcal{V}_Q$ and U are RMSE-type values and B is an SSE-type benefit, so magnitudes are not comparable across rows. Status labels follow Section 3.1.

Study	Correction → scoring	Contrast	Metric	Status; units	Headline [95% CI]
Rayleigh (4.1), ConvNeXt	ψ_0 or ψ_1, $g = 1$ after every prediction → steps 6–12 outside the 12-mode span	exact insertion vs. autonomy	$\mathcal{V}_Q$ (%)	prospective; 35 files, trajectory 0	ψ_0: -1.52% $[-2.42, -0.85\%]$; ψ_1: -1.29% $[-1.80, -0.82\%]$
Rayleigh follow-ups	same, trajectories 1, 3, and 4	same	$\mathcal{V}_Q$ (%)	trajectory 1: test-informed follow-up; trajectory 3: prospective replication; trajectory 4: test-informed, prospectively frozen, outcome-unseen; 35 files each	trajectory 1 (ψ_0, ψ_1): -3.29% $[-5.35, -1.56\%]$, -1.58% $[-2.50, -0.75\%]$; trajectory 3: -0.70% $[-0.93, -0.43\%]$, -0.16% $[-0.22, -0.09\%]$; trajectory 4: -1.10% $[-1.23, -0.93\%]$, -0.28% $[-0.32, -0.23\%]$, but competence failed
Rayleigh spreading (App. C)	stationary regression gain (OI form); damped $g = 0.25$	arm vs. autonomy (exact insertion: rows above)	$\mathcal{V}_Q$ (%)	post hoc (trajectory 0), prospective check (trajectory 3)	gain failed skill check; trajectory 0: -56%, -161%; damped -0.89%, -0.74%
Supernova sweep (4.2), U-Net	90 exact subset and fraction updates of the newest frame → full state, $L = 7$	all-channel vs. hindsight best subset	U (points)	prospective; 16 files, 96 events	0.27 [0.10, 0.43]
Supernova matched repair (4.3)	subset vs. SSE-matched all-channel	$L = 1$; $L = 7$	U (points)	post hoc; 16 files	-0.52 [-0.72, -0.33]; 0.51 [0.39, 0.64]
Acoustic (4.4), classic U-Net	16 pressure modes at 7 events → 13 delayed states	direct carry vs. autonomy	B (points)	prospective; 200 trajectories, 2 files	7.33 [6.82, 7.86]
Acoustic, ConvNeXt U-Net	same	shadow isolation vs. autonomy	B (points)	same	1.04 [0.90, 1.20]
TGC (4.5)	3 channel means in a 24-mode protected span → complement	protected carry vs. emit-only	$\Delta\mathcal{V}_Q$ (points)	prospective; 6 files, 6 modules	0.97 [0.68, 1.36]
Failed routes (4.6)	selector; affine completion; transfer	—	—	post hoc, exploratory	App. D

A.1 EVALUATION ESTIMANDS AND UNCERTAINTY

Metrics by study. Rayleigh–Bénard reports the pooled value $\mathcal{V}_Q$ of Eq. (2) in percent on complement error. TGC computes $\mathcal{V}_Q$ within each fold-seed role and averages the six roles equally, on complement error and on the full state; its carry contrast is $\Delta\mathcal{V}_Q$. Supernova reports the per-event utility U on the full state; and Acoustic reports B in percentage points of autonomous full-state SSE, split into its P and complement parts where stated.

All intervention arms are paired: shared history, checkpoint, and reference attribute within-event differences to the interface. Supernova point estimates give the two mass regimes equal weight. Paired 95% intervals use 20,000 deterministic bootstrap draws, resampling complete files independently within mass; no voxel or event is treated as an independent file. These intervals quantify variation among the available simulations, not generalization across architectures or physical systems.

For TGC, each bootstrap draw keeps both folds, resamples the three seed roles, then resamples files within density strata and trajectories within file. The Acoustic test resamples 100 complete trajectories within each of two files; resampling the 175 distinct static-maze hashes is a dependence sensitivity.

For the Supernova support estimands, let $\mathcal{A}_k$ be all exact subsets of cardinality k. We report the equal-mass mean of $U_{i,A,1,L}$ over events and $A \in \mathcal{A}_k$, so the cardinality curve integrates over channel identity. The hindsight best subset and its mean full-state gap are

$$O_{i,L} = \max_{A \subsetneq \{1,\dots,6\},\, A \neq \varnothing} U_{i,A,1,L}, \qquad D_L = \mathbb{E}_i[U_{i,\mathrm{full},1,L} - O_{i,L}], \tag{11}$$

where the outer expectation gives the two masses equal weight. Because the maximum uses realized future truth, $O_{i,L}$ is a nondeployable upper benchmark for policies restricted to these 62 proper exact subsets at the same horizon and endpoint. A positive D_L means that no selector over that inventory closes the mean full-state gap, even with hindsight; it does not imply pointwise dominance, so we also report event and file incidence.

Fraction controls hold support fixed and change correction amount; nested controls add one channel over every parent-child pair; and the fixed-cardinality comparison tests whether a named semantic grouping explains the trend. None selects the primary horizon or uncertainty unit. For the separately labeled post-hoc matched-repair test, the all-channel fraction is $g_i^A = 1 - \sqrt{E_i^A/E_i^0}$, computed from the historical immediate SSE of each proper exact subset and the raw state. The paired arms therefore remove equal immediate full-state SSE up to a prespecified tolerance before rollout. This matches one scalar repair quantity, not correction direction, information content, norm, or exactness pattern.

A.2 Study-specific protocols

These details complete the compact designs reported with their corresponding results in Section 4. Figure 6 shows one example field from each system.

Rayleigh–Bénard. The counterexample uses the official ConvNeXt U-Net and native trajectory 0 from each of 35 test files. The 12 physical modes are six buoyancy modes and six divergence-free velocity modes defined by a streamfunction basis, in the published order of the fixed implementation. Development on one trajectory from each of 15 validation clusters ranked the 12 exact ($g = 1$) repeated corrections; the development probe recovered the two harmful streamfunction-mean modes in all 15 leave-one-cluster-out analyses, and both were fixed before any test file was downloaded. Every correction is applied exactly ($g = 1$) after every prediction; steps 6 through 12 are scored outside the pressure mean and the complete span, and files are the uncertainty units. Later uses of these test files are listed in Appendix C.

Supernova. Supernova-64 has density, pressure, temperature, and three velocity channels. The official four-frame 64^3 U-Net is evaluated on 16 previously untouched public files: four at $0.1\,M_\odot$ and twelve at $1\,M_\odot$. Each file contributes two trajectories and starts 0, 16, and 32, yielding 96 events. At each event, an intervention replaces channels only in the newest predicted frame; the three earlier history frames remain unchanged. Every event evaluates the 90 interventions in Section 4.2 for seven autonomous steps. The primary comparison at $L = 7$ is all-channel exact insertion against the per-event best of all 62 proper exact subsets; $L \in \{1, 3, 5\}$ are prespecified secondary horizons. Cardinality, nested support, fixed-cardinality coherence, insertion fraction, monotonicity, harm rescue, competence, and physical diagnostics were all prespecified. Missing files or failed arms could not be replaced.

Acoustic Scattering Maze. The dataset supplies official classic and ConvNeXt U-Net checkpoints. Exploration on one 100-trajectory development file fixed the selection rule and chose direct carry for the former and shadow isolation for the latter. A disjoint, untouched 100-trajectory qualification file had to reproduce both directions before final-test access; the final test pooled 200 untouched trajectories from two further public files. Seven exact events at one-indexed steps 4, 8, 12, 16, 20, 24, and 28 insert the joint 16-mode pressure observation. For each trajectory, the 16-mode pressure basis is obtained by masking the 4×4 lowest two-dimensional DCT products to the fluid path and orthonormalizing the 16 masked fields by QR; it is recomputed only when the static maze changes. Interface-emitted states are scored before any same-step observation at steps 17 to 19, 21 to 23, 25 to 27, and 29 to 32, totaling 13 states. Comparators are autonomy, the unselected recurrent interface, event-only complement reset, emit-only correction, and persistence. Choices, checks, data identities, and aggregation were fixed before test access. The two jobs for the first test file were recovered after a manifest-writing error that did not touch their outputs; outcomes were not inspected, the failed and recovered statistics are identical, and the artifact README records the recovery.

Acoustic mechanism checks. Shadow isolation matches the autonomous Q component by construction; the maximum relative Q-SSE mismatch is 2.64×10^{-10}. The P-benefit contrast, shadow isolation minus direct carry, is 0.00 points $[-0.02, 0.02]$ for classic U-Net and 0.05 $[0.03, 0.08]$ for

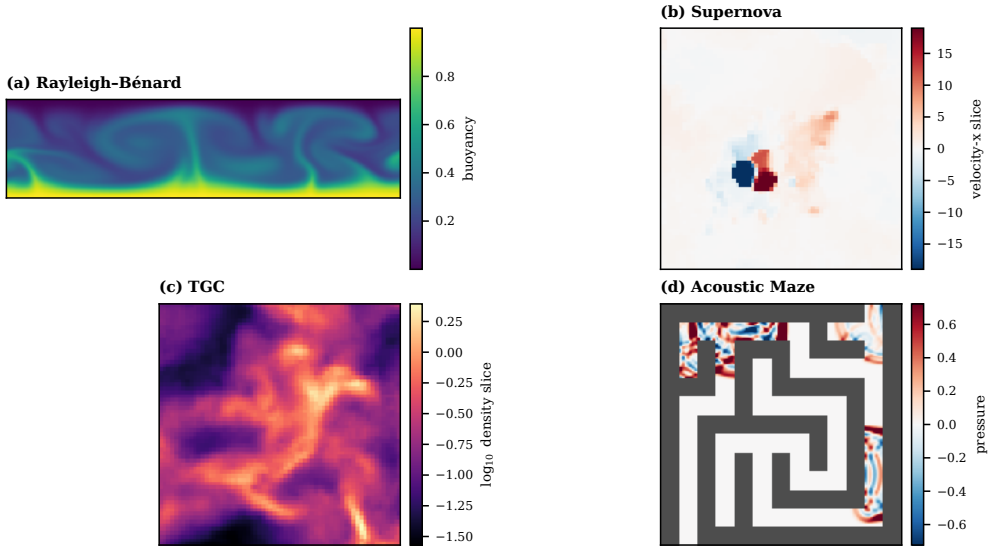

Figure 6: Systems at a glance, shown only for physical context. The fixed examples are Rayleigh–Bénard validation trajectory 0 at step 100; the Supernova event slice used in Figure 7; a mid-plane density slice of TGC validation trajectory 0 at step 25 ($\rho_0 = 4.45\,\mathrm{cm}^{-3}$, $Z = 0.1\,Z_0$ for reference metallicity Z_0, $T_0 = 10\,\mathrm{K}$); and Acoustic development trajectory 0 at step 16, with solid walls in gray. Each panel uses its own color scale and preserves the source domain's spatial aspect ratio. The signed Supernova and Acoustic limits are symmetric 99th-percentile clippings. These images are shown only for physical context and are not additional inferential evidence; the Supernova slice inherits the post-hoc selection disclosed in Figure 7.

Table 2: Prospective acoustic test. Entries are percentage reductions in delayed ratio-of-sums normalized SSE; competence compares autonomy with persistence. Brackets are paired 95% complete-trajectory intervals stratified by test file. The unselected interface is shadow isolation for U-Net and direct carry for ConvNeXt. Opposite prespecified choices improve both checkpoints on untouched test trajectories.

Checkpoint	Selected interface	Competence	vs. autonomy	vs. unselected	vs. reset
Classic U-Net	Direct carry	36.58 [35.86, 37.27]	7.33 [6.82, 7.86]	2.53 [2.14, 2.93]	1.99 [1.62, 2.37]
ConvNeXt U-Net	Shadow isolation	29.93 [29.01, 30.75]	1.04 [0.90, 1.20]	1.23 [0.81, 1.70]	0.66 [0.59, 0.73]

Per-trajectory safety check: the selected interface was worse than autonomy on 2/200 classic U-Net trajectories and 0/200 ConvNeXt U-Net trajectories. Across all 200 classic U-Net trajectories, the positive excess from those two failures totals 0.028% of total autonomous SSE summed across all 200 trajectories.

ConvNeXt U-Net. These are also the isolated-carry parts of a post-hoc emit-only shadow blend, which emits the shadow's complement but keeps direct-carry history and whose complement benefit over autonomy is zero by construction: the blend improves on autonomy by 4.80 points $[4.51, 5.11]$ for classic U-Net and 0.99 points $[0.86, 1.14]$ for ConvNeXt, using the same stratified bootstrap.

Turbulence with gravitational cooling. The public dataset covers 27 parameter sets (three initial temperatures, densities, and metallicities); we pair its public surrogate with a partial analysis operator. Development used disjoint public validation files to fit the compact adapter, protection operators, and all numerical thresholds. The untouched test fixes six public files, two folds, and three seeds. Comparators are persistence, the direct-carry baseline, the emit-only adjustment, protected carry, and parity-shuffled carry. A pure no-analysis autonomous arm was not run. The prespecified checks require the direct-carry baseline to beat persistence, positive protected-carry complement and full-

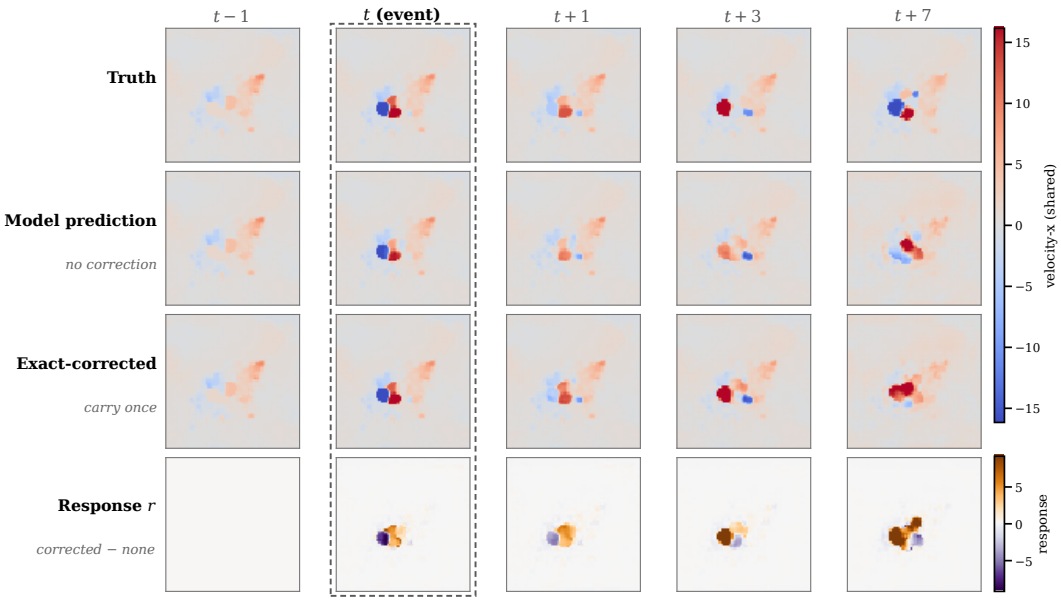

Figure 7: Field anatomy for a post-hoc Supernova rollout. Full-state RMSE is below autonomy at the event and for three steps, then above it from $t + 4$ to $t + 7$. At the dashed event, one exact velocity-x field replaces the newest prediction and is carried without further feedback. The response row shows corrected minus autonomous state. The episode, channel, and plane were selected after viewing outcomes; this is mechanism context, not an additional inferential result.

state value, positive protected-carry minus emit-only complement value, positive effects in all six checkpoint roles, exact protection certificates, and exact replay. No post-test repair or threshold change was allowed.

Provenance. All headline numbers are generated from stored JSON rather than transcribed. The artifact includes compact statistics, controls, runner sources, and aggregation replays that regenerate every reported interval, check, table, and figure; its README files document provenance and execution safeguards. Raw public HDF5 and official checkpoint files are identified by repository revision and integrity metadata but are not redistributed.

For the correction sweep, the protocol, arm inventory, file identities, runner, and aggregator were fixed before test access, files were opened one at a time, and a replay reconstructed all four horizon analyses. The TGC test likewise fixed its files, folds, seeds, arms, and checks before one execution, and a replay reconstructed its bootstrap. The matched-repair test was fixed after the original grid had been accessed but before any new matched rollout; the worst immediate-SSE mismatch is 2.98×10^{-8} relative to raw SSE and 8.28×10^{-6} relative to removed SSE, and a replay reconstructs its summary.

A.3 Secondary Supernova controls

Hindsight-gap sensitivity. The gaps between all-channel exact insertion and the per-event best proper exact subset are 0.98 points at $L = 1$ $[0.71, 1.23]$, 1.23 at $L = 3$ $[0.99, 1.46]$, 0.64 at $L = 5$ $[0.47, 0.80]$, and 0.27 at $L = 7$ $[0.10, 0.43]$. The equal-file sensitivity is 0.21 points, with 12/16 positive file means (one-sided sign test $p = 0.038$). At $L = 7$, all-channel utilities are 9.91% and 10.69% for the 0.1 $M_\odot$ and 1 $M_\odot$ regimes; the corresponding hindsight-best utilities are 9.53% and 10.53%, giving gaps of 0.38 and 0.15 points.

Matched-repair sensitivity. Averaged over all 62 proper subsets, channel-restricted-exact minus fraction-scaled matched-all-channel utility is -0.52 points at $L = 1$ $[-0.72, -0.33]$, 0.22 at $L = 3$ $[0.07, 0.35]$, and 0.51 at $L = 7$ $[0.39, 0.64]$. The sign reversal persists under every leave-one-file-out analysis: the $L = 1$ range is $[-0.574, -0.462]$ points and the $L = 7$ range is $[0.479, 0.549]$. A per-event linear benchmark, calibrated by all-channel insertion, makes a correction's delayed SSE

Table 3: Post-hoc Supernova matched-repair test at $L = 7$. Values are exact-subset minus actual matched-all-channel RMSE utility in percentage points. Intervals resample the 16 complete files within mass (4 and 12 files). The prespecified centered-bootstrap test is at its resolution limit (raw $p = 1/20001$), giving Holm-adjusted $p = 2.5 \times 10^{-4}$ for all five. A post-hoc exact file-level sign-flip test over all 2^{16} sign vectors gives Holm-adjusted $p \approx 0.002$ for all five.

Exact channels	Utility difference (points)	Holm p
1	0.20 $[0.13, 0.26]$	2.5×10^{-4}
2	0.40 $[0.29, 0.50]$	2.5×10^{-4}
3	0.56 $[0.42, 0.70]$	2.5×10^{-4}
4	0.66 $[0.49, 0.82]$	2.5×10^{-4}
5	0.60 $[0.45, 0.75]$	2.5×10^{-4}

reduction proportional to the fraction of immediate SSE it removes. It accounts for most of the raw cardinality trend, and exact subsets deliver less than their proportional share: every complete-file interval of the residual, actual minus proportional reduction, is below zero. At $L = 7$, actual minus interpolated matched-all-channel SSE reduction is -0.11 points $[-0.13, -0.09]$.

The post-hoc comparator uses file indices 01 to 03 in each mass regime, fixed before computing its outcomes, and distributes the event-time full-state increment over four recurrent states. With six files, a file-level sign-flip test cannot go below $p = 1/32$, so its intervals are descriptive. Table 4 gives its full horizon profile. Although distributing the increment achieves 6.68% value at $L = 7$, compared with 3.61% for its paired one-shot $g = 0.25$ arm on this same six-file cohort, a difference of 3.07 points $[2.56, 3.90]$, it remains below one-shot exact insertion at every scored horizon. The IAU-minus-exact contrast ranges from -5.25 to -1.93 points across the six files and from -4.16 to -3.08 under leave-one-file-out analysis. One of 36 IAU-style episodes is harmful relative to autonomy (2.78%), with positive excess equal to 0.008% of total autonomous SSE; both one-shot arms have zero harmful episodes. A single harmful episode is not interpreted as a population-level difference.

Separately, on the 16-file prospective correction sweep, at $L = 7$ all-channel newest-frame fractions $g = 0.25, 0.5, 0.75, 1$ yield 3.41%, 6.48%, 8.82%, and 10.30% utility, respectively, with zero harmful events at every gain. Adjacent increments are 3.07 points $[2.96, 3.19]$, 2.34 points $[2.23, 2.48]$, and 1.48 points $[1.30, 1.65]$. A larger fraction helps on this fixed grid; an untested continuous fraction could still outperform $g = 1$.

Adding an exact channel is not uniformly beneficial. Across all nested pairs, the improving fractions for $1 \to 2$, $2 \to 3$, $3 \to 4$, $4 \to 5$, and $5 \to 6$ channels are 89.51%, 91.55%, 92.63%, 93.87%, and 95.60%. At fixed size three, the thermodynamic block exceeds the other subsets by 0.69 points and lies at the mean 60.64th event-wise percentile. Channel identity matters, but it does not explain the larger raw cardinality trend.

B PROTECTED TGC IMPLEMENTATION

The normalized two-state history is flattened to 12 channels and passed through a trainable $3 \times 3 \times 3$ convolution into the frozen U-Net input stem. The modified-minus-direct frozen-host output defines the raw learned adjustment R. The protected span is a 24-mode orthonormal basis (six channels, each with its constant and first cosine mode in x, y, and z), of which the analysis observes three channel means. The forecast-anchored projection Π_d removes the adjustment's protected-span component before emission, giving $r_t = \Pi_{d_t}(R(h_t))$, and outer constraints keep positive physical fields positive. For recurrence, Π_A keeps the analysis increment only in the parity sectors that are odd in at least two axes (a reflection-closed subspace that excludes every protected mode); the increment is scaled to respect the outer physical constraints and cast to host precision. A certificate checks protected-mode leakage, analysis residual, finite values, no new nonpositive voxel, and bitwise identity of the consumed state; no fallback occurs in the test.

The implementation of A_x, denoted C_{32} in the code and certificates, repeatedly removes residuals in the active orthonormal analysis coordinates on the actual float32-consumed tensor. It stops once

Table 4: Post-hoc Supernova IAU-style fixed-increment insertion. Values are RMSE-type value over autonomy in percent; the final column is IAU minus exact one-shot value, in points. Intervals are descriptive 95% complete-file bootstrap intervals over three files per mass regime. Equal nominal vector sums do not match achieved immediate removal, correction energy, or realized trajectory dose.

Horizon L	IAU-style, $K = 4$	Exact one-shot	Difference (points)
1	13.81 $[12.33, 15.37]$	30.56 $[26.17, 31.97]$	-16.76 $[-18.72, -13.58]$
3	9.78 $[8.87, 10.72]$	19.12 $[17.77, 19.65]$	-9.33 $[-10.46, -8.03]$
5	8.00 $[7.39, 9.29]$	13.56 $[12.91, 14.25]$	-5.56 $[-6.47, -4.42]$
7	6.68 $[6.02, 7.74]$	10.28 $[9.56, 11.05]$	-3.60 $[-4.45, -2.61]$

the absolute residual is at most 10^{-8} or the relative residual is at most 10^{-6}, when float32 rounding prevents further reduction, or after 32 monotone refinements; its certificate requires the stated residual tolerance.

Each development fold uses three files, five trajectories per file, two fixed passes, and seeds 17, 23, and 41. The loss is emitted complementary-subspace error relative to the direct-carry history. Neither test files nor future test targets tune the adjustment scale or projection. Each of the six fixed fold-by-seed modules is applied to all six reserved files; effects are computed within module before equal averaging, so modules are procedure replicates, not additional independent data.

C Rayleigh uncentered-regression spreading baseline

Status and prior use of the test files. The protocol, runner, and interpretation rules were frozen before any rollout of this study, and the validation-estimated gains and selections were frozen before any test tensor was read. Table 5 lists every use of the 35 Rayleigh–Bénard test files. Ordinals 1 and 2 were test-informed follow-ups, both with failed primary claims; apart from the ordinal-1 harm replication cited in Section 4.1, they are not reported elsewhere. Ordinal 3 had never been decoded before this study (its bytes had only been checksummed), so it is prospective. It comes from the same 35 $(\mathrm{Ra}, \mathrm{Pr})$ cells, so it is a same-cell replication, not an independent sample. Every test trajectory used here (ordinals 0 and 3) shares its first state with a validation trajectory of another cell and none with one of its own cell, so no trajectory is duplicated, but no overlap-free test subset exists.

Table 5: Uses of the Rayleigh–Bénard test files, by native trajectory ordinal.

Ordinal	Study	Outcome
0	Signed-probe test (Section 4.1)	ranking primary failed; both harms replicated
1	Three-step confirmation	primary failed ($\rho = 0.678$); harms -3.29% $[-5.35, -1.56\%]$, -1.58% $[-2.50, -0.75\%]$
2	Selective-policy test	primary failed (selective vs. full $[-0.05, 5.27\%]$)
0, 3	This study	Table 7
4	Frozen competence-gated check	both complement harms recur, but the confirmation fails because autonomy is worse than persistence (Table 6)

Failed competence-gated confirmation. After the trajectory-0, 1, 2, and 3 analyses, we froze one final check on the outcome-unseen ordinal 4 using the exact original 26-branch batch, the same 35 files, checkpoint, two selected modes, scoring window, and whole-file bootstrap. Confirmation required both negative complement intervals and an autonomous competence interval above zero. The complement condition holds for both modes, and each selected mode is also worse than full-12 insertion, but competence fails sharply (Table 6). We therefore treat this only as evidence that the partial-update response persists in an incompetent rollout regime, not as another competent-simulator confirmation. The signed probe also fails its conjunctive primary: its top-two set matches the repeated-correction top two, but both miss the frozen validation pair, despite $\rho = 0.769$ ($[0.706, 0.776]$). Complement harm does not imply full-state harm here: ψ_0 improves full-state error, while ψ_1 is not detectably different from autonomy.

Table 6: Outcome-unseen Rayleigh trajectory-4 check. Values are RMSE utility in points; positive is better. Intervals resample the 35 complete files. The frozen confirmation fails because the autonomy-versus-persistence competence interval is below zero. The all-12 interval crossing zero means no detectable complement effect, not safety.

Contrast	Value	95% interval
Autonomy vs. persistence	-103.62%	[-238.06, -47.83%]
Exact ψ_0 vs. autonomy (complement)	-1.10%	[-1.23, -0.93%]
Exact ψ_1 vs. autonomy (complement)	-0.28%	[-0.32, -0.23%]
Exact all-12 vs. autonomy (complement)	0.59%	[-0.12, 1.53%]
Exact ψ_0 minus all-12 (complement)	-1.69%	[-2.51, -1.08%]
Exact ψ_1 minus all-12 (complement)	-0.87%	[-1.77, -0.19%]
Exact ψ_0 vs. autonomy (full state)	3.023%	[2.945, 3.102%]
Exact ψ_1 vs. autonomy (full state)	0.0058%	[-0.0005, 0.0134%]

Design. After prediction t with truth error e_t, exact insertion adds $d_t\phi_k$, where $d_t = \langle e_t, \phi_k \rangle$ is the exact coefficient of the corrected mode. The spreading arm adds $d_t(\phi_k + \alpha K)$. The primary gain $K_0 = \sum d\, r / \sum d^2$ regresses the uncorrected error $r = e - d\phi_k$ (pressure mean removed, orthogonal to ϕ_k) on d; it is the least-squares gain for the applied update. It coincides with the centered covariance/OI gain $C_{rk}C_{kk}^{-1}$ when the relevant errors are mean-zero; equality can also occur under other mean configurations. It is estimated from true errors, a best case for estimating background statistics, after predictions 1–11 of all 175 validation trajectories (five per file) in rollouts with repeated exact insertion, pooled over lead times and cells. The damped arm adds $\gamma d_t \phi_k$. Exploratory variants split K_0 into its complement and 12-mode-span parts, average it along the periodic direction, center it, re-estimate it from its own cycle (K_1), and correct both modes jointly. Shrinkage $\alpha^\star \in \{0, 0.5, 1\}$ was selected on validation with gains cross-fitted over the 50 groups of trajectories that share a first state, and $\gamma^\star \in \{0.25, 0.5, 0.75, 1\}$ on validation (γ is the insertion fraction g). The gain was declared to have skill only if its cross-fitted held-out R^2 for the complement error was positive and, cycled on validation, it lowered the post-analysis complement error summed over the first eleven analyses relative to exact insertion. The two harmful modes were selected on 15 of these validation trajectories, which biases their validation harm. The sealed branches of 15 development trajectories and of all 35 ordinal-0 test files reproduced the sealed results bit for bit, and ordinal 3 was decoded only after all 35 ordinal-0 files passed. The utility $\mathcal{V}_Q$ is that of Section 4.1.

Validation. With cross-fitted gains, the validation values are -1.60% (exact insertion) and -434% (spreading, $\alpha = 1$) for ψ_0, and -1.13% and -108% for ψ_1, so validation selects no spreading ($\alpha^\star = 0$). Damping reduces the harm monotonically; the selected $\gamma^\star = 0.25$ gives -0.92% and -0.64% and is the smallest grid value, so validation favors little insertion. The gain is stable across folds (mean spread 0.01% and 0.40% of its energy) and has positive held-out R^2 (0.24 and 0.10), but, cycled, it raises the summed post-analysis complement error to 32 and 4.8 times that of exact insertion, so the skill check fails for both modes. On test ordinal 0 the same quantities are $R^2 = 0.04$ and 0.05 and ratios 2.7 and 7.6, and on ordinal 3 $R^2 = 0.60$ and 0.20 and ratios 84 and 1.4, in agreement. The ψ_0 coefficient is strongly biased (mean over standard deviation -2.34; ψ_1: 0.10), so for ψ_0 the uncentered gain is a regression gain rather than a covariance gain, and the centered variant also worsens (Table 7).

A post-hoc descriptive analysis, not prespecified, locates the failure. At the first analysis, where both arms start from the same prediction, spreading multiplies the complement error by 11 and 5.2. The least-squares scale of the pooled complement gain fitted at each lead alone rises from 0.19 at prediction 1 to 1.79 at prediction 11 for ψ_0 (from -0.04 to 1.38 for ψ_1), so the pooled gain is far too large early and too small late; for ψ_1 its direction also differs at the first predictions. The pooled held-out R^2 is nonetheless positive because it is dominated by late leads, where the exact-insertion complement error is 142 and 140 times larger than at prediction 1. We read this as an early overshoot that compounds in the cycle; consistent with it, the gain re-estimated from that cycle points the other way (cosine -0.88 and -0.63 with K_0).

Test. Table 7 reports every arm on both ordinals. By the frozen rules, all four primary cells fall in "worsens" or "not detectable", and the gain lacks skill for both modes. Applied, the gain deepens the

harm on trajectory 0 (-56% and -161%); trajectory 3 agrees for ψ_0 (-790%), and for ψ_1 (-19%) the change from exact insertion is not detectable. The validation-selected damped arm ($\gamma^\star = 0.25$) gives -0.89% and -0.74% on trajectory 0, and every damped arm reduces but does not remove the harm, on both ordinals. Of six exploratory variants (unadjusted), one reaches "helps" on trajectory 0 (ψ_1, spreading only within the 12-mode span) and does not replicate on trajectory 3. Because the gain fails its skill check and its applied values reflect an unstable cycle (above), we withdraw it as evidence about covariance spreading; the main text reports only the replication and the damped arm.

Scope. The comparison tests one stationary gain pooled over lead times and the 35 cells, estimated from errors cycled with exact insertion. Flow-dependent, ensemble, lead-time-dependent, per-regime, cycle-consistent (iterated), and bias-aware gains are untested. Supernova is excluded because its regression-gain comparator is a failed development-only screen (Section 4.6), and Acoustic because its per-maze 16-mode basis makes a pooled modal covariance ill-posed. Departures from the originally planned design: the all-12-mode spreading arm was dropped because exact all-12 correction is not harmful, localization was replaced by the periodic average and cross-fitted skill diagnostics, shrinkage was selected leave-one-first-state-out, and no test-fitted gain was run.

Deviations. The protocol directory lists twelve departures from the frozen protocol. One is analytical: after test access, the frozen aggregator crashed on a descriptive step, because every test trajectory used shares its first state with a validation trajectory of another cell and the prescribed re-aggregation without overlapping files therefore had no files left; a separate script runs the frozen code unchanged, records that step as undefined, and was committed before any test value was aggregated. The other eleven concern manuscript wording and display: the skill-check phrasing; the headline wording and a factual clause on trajectory 0, and the later removal of that headline sentence from the abstract, Related Work, and conclusion; the note that the best exploratory variant does not replicate; the trajectory-3 replication statements; the length of the Section 4.1 paragraph; minor template wording; the display of these results in Figure 2 and Table 1; and the renaming of the frozen "covariance spreading" and optimal-interpolation gain to the uncentered regression gain actually fitted; and moving the spreading, damped-gain, and exploratory results from Section 4.1 to this appendix. None changes an arm, estimand, category, skill verdict, or headline bin.

D FAILED REMEDIES AND TRANSFER SCREENS

Crossing the original rollout-based deficit screen with eight matched scorers shows that response-only and the full critic make identical decisions. Response-only minus correction-only is -0.097 points $[-0.204, 0.000]$, response-only minus channel-only is -0.003 $[-0.009, 0.000]$, and the screen lowers response-only utility by 0.838 points $[0.149, 1.615]$.

Affine completion is a ridge regression that spreads the revealed residual into the unobserved channels at each voxel, with no spatial spreading. Its observation-only arm is the closest analogue of optimal interpolation; the observation-aware arm also uses forecast and previous-state features. On four development files, observation-aware affine completion improves seven-step utility over observed-only insertion (3.430% versus 2.542%) but is slightly worse than observation-free state steering (3.471%). It recovers 9.923% of the all-channel exact opportunity and violates its prespecified immediate- and horizon-1 harm checks. On Shear Flow, emit-only and protected-carry values are -7.738% and -24.365%; on three Viscoelastic files they are 0.903% and 0.841%, with a -0.062-point pooled carry contrast. These were closed as negative routes rather than promoted.

A separately prespecified retrospective diagnostic on the previously accessed Acoustic development cohort used the negative VJP of delayed complement SSE with respect to 16 shared correction-mode gains at zero gain. It failed to rank finite-dose repeated-correction utility: Spearman ρ was -0.079 for the classic U-Net and 0.168 for ConvNeXt. It was worse than the current-error proxy for classic U-Net and had no supported advantage for ConvNeXt; the combined prespecified check failed. No Acoustic test cohort was opened for this diagnostic. This closes that local signed-gradient ranking surrogate.

We next froze a distinct retrospective correction-gain Jacobian diagnostic on the same development cohort. For the exact gain-one response $R = e(1) - e(0)$, the normalized complement-SSE benefit decomposes identically as $-2\langle e(0), R\rangle / \|e(0)\|^2 - \|R\|^2/\|e(0)\|^2$. Table 8 shows that signed align-

Table 7: Rayleigh uncentered-regression spreading test results (the frozen study name used "covariance spreading"): pooled complement value $\mathcal{V}_Q$ on predictions 6–12, contrast Δ with exact insertion of the same mode (points), 95% file-bootstrap intervals conditional on the frozen gains, and the prespecified category. The primary arm and the selected damped gain are prespecified; grid rows are descriptive; the remaining variants are exploratory and unadjusted. Validation selected $\alpha^\star = 0$, i.e. the exact-insertion arm, whose category is therefore "not detectable" ($\Delta = 0$).

Arm	$\mathcal{V}_Q$	interval	Δ	interval	category
Trajectory 0 (post hoc)					
ψ_0 exact insertion	−1.52%	[−2.42, −0.85]	—	—	—
OI, K_0, $\alpha = 1$ (primary)	−56.10%	[−88.25, −37.60]	−54.58	[−86.44, −36.33]	worsens
OI, K_0, $\alpha = 0.5$ (grid)	−10.38%	[−18.44, −5.50]	−8.86	[−16.27, −4.36]	worsens
Damped, $\gamma = 0.25$ (selected)	−0.89%	[−1.44, −0.48]	0.63	[0.37, 0.99]	reduces, harmful
Damped, $\gamma = 0.5$ (grid)	−1.24%	[−2.00, −0.67]	0.28	[0.17, 0.43]	reduces, harmful
Damped, $\gamma = 0.75$ (grid)	−1.42%	[−2.27, −0.78]	0.10	[0.07, 0.16]	reduces, harmful
OI, complement part K_0^Q	−69.07%	[−107, −47.52]	−67.54	[−105, −46.21]	worsens
OI, span part K_0^P	−0.78%	[−2.13, 0.14]	0.75	[−0.42, 1.73]	not detectable
OI, x-averaged $K_0^{\bar{x}}$	−42.50%	[−66.45, −29.03]	−40.97	[−64.45, −27.77]	worsens
OI, centered K_c	−104%	[−155, −75.32]	−103	[−152, −74.04]	worsens
OI, second iterate K_1	−208%	[−291, −162]	−207	[−289, −160]	worsens
ψ_1 exact insertion	−1.29%	[−1.80, −0.82]	—	—	—
OI, K_0, $\alpha = 1$ (primary)	−161%	[−218, −128]	−160	[−217, −127]	worsens
OI, K_0, $\alpha = 0.5$ (grid)	−38.51%	[−66.35, −22.28]	−37.22	[−65.10, −20.99]	worsens
Damped, $\gamma = 0.25$ (selected)	−0.74%	[−1.04, −0.45]	0.56	[0.37, 0.77]	reduces, harmful
Damped, $\gamma = 0.5$ (grid)	−1.04%	[−1.46, −0.65]	0.25	[0.18, 0.34]	reduces, harmful
Damped, $\gamma = 0.75$ (grid)	−1.20%	[−1.68, −0.76]	0.09	[0.07, 0.12]	reduces, harmful
OI, complement part K_0^Q	−165%	[−227, −130]	−164	[−225, −129]	worsens
OI, span part K_0^P	4.36%	[1.50, 6.24]	5.66	[2.60, 7.78]	helps
OI, x-averaged $K_0^{\bar{x}}$	−35.90%	[−74.75, −12.15]	−34.60	[−73.43, −10.80]	worsens
OI, centered K_c	−149%	[−199, −116]	−147	[−198, −115]	worsens
OI, second iterate K_1	−2536%	[−3338, −2073]	−2534	[−3336, −2072]	worsens
Joint exact insertion (ψ_0, ψ_1)	−2.41%	[−3.67, −1.46]	—	—	—
Joint OI G_J	−349%	[−466, −284]	−346	[−463, −282]	worsens
Trajectory 3 (prospective)					
ψ_0 exact insertion	−0.70%	[−0.93, −0.43]	—	—	—
OI, K_0, $\alpha = 1$ (primary)	−790%	[−840, −725]	−790	[−839, −724]	worsens
OI, K_0, $\alpha = 0.5$ (grid)	−236%	[−258, −207]	−235	[−257, −207]	worsens
Damped, $\gamma = 0.25$ (selected)	−0.36%	[−0.48, −0.21]	0.34	[0.21, 0.46]	reduces, harmful
Damped, $\gamma = 0.5$ (grid)	−0.55%	[−0.73, −0.33]	0.15	[0.09, 0.21]	reduces, harmful
Damped, $\gamma = 0.75$ (grid)	−0.64%	[−0.86, −0.39]	0.06	[0.03, 0.08]	reduces, harmful
OI, complement part K_0^Q	−759%	[−800, −706]	−759	[−799, −706]	worsens
OI, span part K_0^P	−89.68%	[−98.17, −78.75]	−88.98	[−97.29, −78.25]	worsens
OI, x-averaged $K_0^{\bar{x}}$	−431%	[−455, −400]	−431	[−454, −399]	worsens
OI, centered K_c	−529%	[−554, −503]	−528	[−554, −502]	worsens
OI, second iterate K_1	−664%	[−706, −625]	−663	[−705, −624]	worsens
ψ_1 exact insertion	−0.16%	[−0.22, −0.09]	—	—	—
OI, K_0, $\alpha = 1$ (primary)	−18.76%	[−49.20, 4.57]	−18.60	[−49.04, 4.76]	not detectable
OI, K_0, $\alpha = 0.5$ (grid)	7.10%	[−8.08, 19.09]	7.26	[−7.94, 19.28]	not detectable
Damped, $\gamma = 0.25$ (selected)	−0.08%	[−0.11, −0.04]	0.08	[0.04, 0.11]	reduces, harmful
Damped, $\gamma = 0.5$ (grid)	−0.12%	[−0.17, −0.07]	0.04	[0.02, 0.05]	reduces, harmful
Damped, $\gamma = 0.75$ (grid)	−0.14%	[−0.20, −0.08]	0.02	[0.01, 0.02]	reduces, harmful
OI, complement part K_0^Q	−36.69%	[−65.03, −15.59]	−36.53	[−64.90, −15.42]	worsens
OI, span part K_0^P	−0.20%	[−0.82, 0.34]	−0.04	[−0.67, 0.49]	not detectable
OI, x-averaged $K_0^{\bar{x}}$	−3.12%	[−23.06, 11.56]	−2.96	[−22.92, 11.76]	not detectable
OI, centered K_c	−12.17%	[−40.85, 9.95]	−12.01	[−40.70, 10.14]	not detectable
OI, second iterate K_1	−2381%	[−3126, −1557]	−2381	[−3126, −1557]	worsens
Joint exact insertion (ψ_0, ψ_1)	−0.86%	[−1.16, −0.53]	—	—	—
Joint OI G_J	−837%	[−868, −796]	−836	[−867, −795]	worsens

ment exceeds response energy for the classic U-Net but not for ConvNeXt. This post-hoc identity describes the realized response; it is neither a predictor nor an architecture-level explanation.

Table 8: Post-hoc exact response decomposition on the already accessed Acoustic development cohort. All entries are percentages of autonomous ($g = 0$) complement SSE. The final column gives paired static-maze cluster-bootstrap stability intervals.

Checkpoint	Signed alignment	Response energy	Net benefit [95% interval]
Classic U-Net	13.06	10.34	$+2.72$ $[2.08, 3.33]$
ConvNeXt U-Net	11.05	12.51	-1.46 $[-2.17, -0.84]$

The separately prespecified local diagnostic estimated all 16 mode-gain columns by central differences at $\epsilon = 10^{-3}$, with each gain shared across the seven events. It failed before interpretation: forward and backward one-sided derivatives disagreed by relative ℓ_2 values 2.69 and 1.82 for the classic and ConvNeXt checkpoints. Relative ℓ_2 discrepancies between the local and exact gain-one responses were 146.21 and 13.41, with cosines 0.001 and 0.015. The emitted-but-not-carried null was exactly zero, and the exact decomposition closed to numerical precision, but these controls do not repair the failed local-validity checks. We therefore make no local-Jacobian mechanism claim and do not interpret its unstable norm or checkpoint contrast.

A subsequent retrospective forward-mode autodiff JVP audit on the same development cohort also failed its frozen gates: the classic checkpoint exceeded the JVP-additivity tolerance, the ConvNeXt checkpoint exceeded the no-carry-null tolerance, and both failed the local-to-gain-one compatibility check. Because max pooling uses PyTorch's selected subgradient at activation boundaries, and the numerical-validity gate failed, we do not interpret the JVP magnitudes; this audit does not rescue the finite-difference study.

E Response-selector details

The target-free counterfactual response critic (TCRC) was the original selector route. The matched controls in Appendix D falsified its claimed architectural advantage; we retain these details only so the negative result is auditable.

E.1 Feature inventory

Every candidate has 25 target-free scalar features: the correction's log norm, relative norms, and tendency cosine; lag-1/final response gain, gain slope and mean, and relative response energy; minimum/mean temporal cosines with the preceding response, event tendency, and autonomous increment; maximum/mean response-to-autonomous ratios and accumulated cosine/work ratio; and, for density, pressure, and temperature, carry-minus-discard nonpositive fraction and log positivity-deficit ratio. No post-event reference enters a feature.

The input additionally includes a six-dimensional one-hot channel identity. Log features clamp positive arguments at 10^{-30}. The final regressor uses scikit-learn's deterministic gradient boosting implementation with 100 trees, learning rate 0.03, depth 2, minimum leaf size 24, Huber loss, and one seed. Selection receives a sanitized object containing only channel identity, features, and physical diagnostics; reference-dependent SSEs and labels are removed before the policy is called. The outcome aggregator reconstructs this view and requires exact prediction, eligibility, and action replay.

F Rayleigh controls and excluded exploratory evidence

Rayleigh counterexample details. On all 35 held-out files, the Rayleigh checkpoint improves over persistence by 40.15% ($[27.13, 52.62\%]$). Nevertheless, each of the two streamfunction-mean corrections selected as harmful on validation is harmful on test: -1.52% ($[-2.42, -0.85\%]$) for ψ_0 and -1.29% ($[-1.80, -0.82\%]$) for ψ_1. Each is harmful on $21/35$ files. Their full-state values are -0.96% and -0.93%, respectively. Full 12-mode correction has value -0.06% ($[-1.38, 2.15\%]$); the interval crosses zero. A one-step signed probe identifies the harmful pair, but its frozen general-

ranking claim fails: $0.594 < 0.70$ ($[0.510, 0.951]$), with only $7/12$ signs correct. Its point correlation exceeded those of modal-error and response-energy baselines.

Excluded MHD exploration. We do not use an earlier MHD exploration as evidence. The public checkpoint's output convention could not be established from its model card, configuration, or stored metadata, and alternative absolute-state and increment interpretations did not reproduce a competent forecasting interface. The MHD numbers, tables, and figures were therefore removed rather than choosing the more favorable convention after observing outcomes.

G Insertion fraction, horizon, and least-MSE details

Proof of the proposition. If the cross block vanishes, Equation (4) reduces to $p^\top W p \geq 0$. Conversely, if the cross block is nonzero, some $p \in \mathrm{range}(P)$ and $q \in \ker(P)$ have $p^\top W q \neq 0$. Rescaling q makes the mixed term dominate the nonnegative quadratic term, producing harm. □

Zero-intercept least-MSE update. With $e_0 = p + q$, uncentered moments S_{pp}, S_{qp}, and $K = S_{qp} S_{pp}^{+}$, the update leaves $e_1 = q - Kp$ and removes $\delta = (I + K)p$ (written $[I; K]p$). Since $\mathbb{E}[e_1 p^\top] = S_{qp} - K S_{pp} = 0$ (the range of S_{pq} lies in that of S_{pp}) and δ is linear in p, the cross term $2\mathbb{E}[\delta^\top W e_1]$ vanishes, so $\mathbb{E}[e_0^\top W e_0 - e_1^\top W e_1] = \mathbb{E}[\delta^\top W \delta] = \mathrm{tr}(W[I; K] S_{pp} [I; K]^\top) \geq 0$. Subtracting Equation (6) gives the nonnegative difference $\mathrm{tr}(W_{qq} K S_{pp} K^\top) \geq 0$. K is the zero-intercept linear least-MSE gain; by the orthogonality principle it minimizes $\mathbb{E}[e_1^\top W e_1]$ for every PSD W at once. The argument fails with model bias and when S is not the second moment of the errors at that correction time, and it does not cover noisy observations. □

Insertion fraction. The linear calculation of Section 2 separates support from correction amount. For an insertion fraction $g \in [0, 1]$, the corrected current error is $(1 - g)p + q$, so the reduction in finite-horizon squared error is exactly

$$I(g) = 2g\, p^\top W (p + q) - g^2 p^\top W p. \tag{12}$$

When $p^\top W p > 0$, the best continuous fraction is the clipping to $[0, 1]$ of $p^\top W (p + q)/(p^\top W p)$. A positive infinitesimal update requires $p^\top W (p + q) > 0$, whereas the exact update requires the stronger condition $2p^\top W (p + q) > p^\top W p$. Thus a smaller fraction is not necessarily *better*. Within this linear model, if the exact update helps then every smaller positive fraction also helps; however, an infinitesimal update can help while the exact update hurts when only the weaker first condition holds. This motivates evaluating prespecified fractions rather than selecting one after test outcomes.

In Equation (7), even when P and Q select disjoint semantic channels at the event, dynamics can make $P W_L (I - P)$ nonzero by transporting error across channels. For a nonlinear simulator, replacing M^ℓ by the rollout Jacobian gives a local diagnostic but not a certificate: model bias and higher-order terms are absorbed neither by the Jacobian nor by Equation (7). The Supernova sweep consequently reports every prespecified horizon rather than using a one-step sensitivity as a surrogate for delayed value.

Model bias. Suppose the model propagates error with $\hat{M}$ and let b_ℓ be the error at lag ℓ of the model rolled out from the exact current state. The future error is then $\hat{M}^\ell e_0 + b_\ell$, and exact removal of p improves the scored squared error by

$$p^\top \hat{W} p + 2p^\top \hat{W} q + 2p^\top c, \qquad c = \sum_{\ell=1}^{L} (\hat{M}^\ell)^\top Q^\top Q\, b_\ell, \tag{13}$$

with $\hat{W}$ the Gramian of $\hat{M}$. Even when $P\hat{W}(I - P) = 0$, choosing $p = -\epsilon P c$ with small $\epsilon > 0$ gives harm whenever $Pc \neq 0$, so the proposition's condition is not sufficient for a biased model.

Two special cases sharpen the interpretation. If the scored dynamics are block diagonal with respect to P, Equation (5) holds and exact partial correction cannot remove beneficial cancellation across the boundary. If $P = I$ and the reference follows the same exact dynamics, complete state replacement sets current error to zero and is automatically compatible. Neither premise holds for a misspecified learned simulator evaluated against an external trajectory.