AUXILIARY LOCATION INFERENCE FOR RECURRENT SPARSE FIELD RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse field reconstruction with uncertain sensor locations must distinguish prediction errors from spatial misalignment, often with only some vector components observed. We develop calibrated auxiliary location inference with sensor-block cross-fitting and incorporate its probabilities consistently into a kernel correction. On three cylinder cases, the recurrent implementation reduces aggregate NRMSE by 3.40–5.25% over the original auxiliary procedure, retaining 1.63–2.29% in a second observation batch. Matched reference comparisons and component ablations separate the roles of the auxiliary field, likelihood adjustment and stage assignment. We then connect location probabilities to the implemented field response through the observation operator, working covariance, solve and postprocessing. Controlled interventions identify two breaks in this link: physical closure can reverse a raw-field gain, and held-sensor scores can favor corrections that increase full-field error. Cross-physics evaluations show how the measured gains depend on the scene and reference quality.

1 INTRODUCTION

Sparse sensors constrain only a small part of a physical field. Reconstruction therefore depends on a predictive reference, especially when some vector components are missing. If sensor locations are uncertain, a discrepancy may reflect a field error, a location error, or both. Resolving this ambiguity matters twice in a recurrent system: the inferred locations affect the current correction, and the completed field becomes input to the next prediction.

We estimate locations using a low-dimensional auxiliary correction around the reference. Fitting this auxiliary field to a sensor can make that sensor's own location likelihood overly favorable. Sensor-block cross-fitting limits this reuse, while training-fitted calibration adjusts the visible residuals' bias and scale. The resulting location probabilities enter the observation operator, kernel matrix and working covariance together. This couples location inference to a full-field correction supported by both the background basis and the sensors.

The same location estimator also participates in candidate selection. Candidates are fitted on sensor subsets and scored on held sensors; the final correction uses all observations. Our AB implementation uses the original auxiliary estimator A for reference and solver selection, then the cross-fitted, calibrated estimator B for final correction. AB was selected on the earlier development panel and frozen for subsequent comparisons (Appendix T); AA and BB test whether that assignment transfers.

We develop calibrated auxiliary location inference for missing-component, recurrent reconstruction, with sensor-block cross-fitting and consistent probability-weighted kernel correction. Matched reference alternatives and repeated ablations quantify when the auxiliary field and likelihood adjustment add value. We also identify where better location probabilities cease to imply better fields: changes to the observation operator and working covariance alter the correction, closure changes the error geometry, and sparse scores can misrank full-field risks even at true locations. Exact response identities and controlled interventions distinguish these paths; independent rollouts test the resulting algorithm.

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

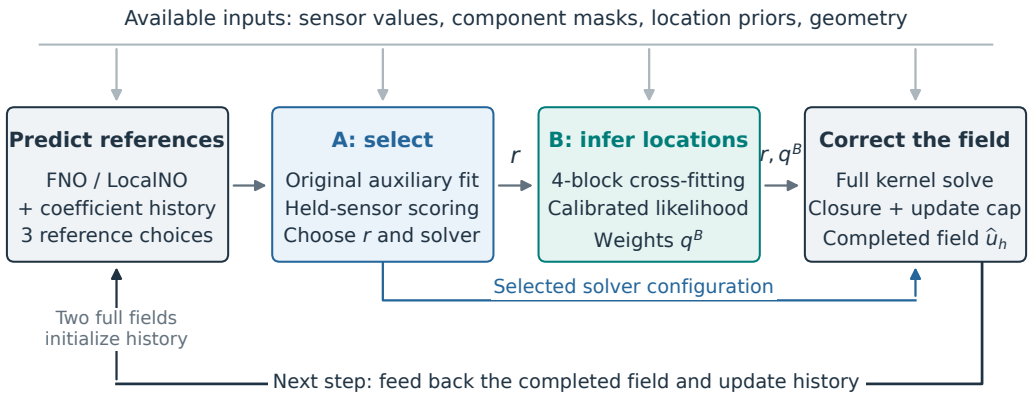

Figure 1: One recurrent step of AB. A selects a reference and solver using held-sensor scores. B recomputes location weights for the final fit with all observations, using four auxiliary fits and training-fitted calibration. The selected solver is retained. Its completed output drives the next prediction; the auxiliary field is used only to interpret locations.

2 Related Work

Gappy POD and PBDW establish reconstruction from incomplete observations and background spaces with observation-supported updates (Everson & Sirovich, 1995; Maday et al., 2015; Maday & Taddei, 2019; Gong et al., 2019). FNO and LocalNO supply our predictive references (Li et al., 2021; Liu-Schiaffini et al., 2024). RecFNO and FLRONet address sparse field reconstruction with neural operators (Zhao et al., 2024; Dang & Nguyen, 2025); our public-architecture comparisons identify their task adapters separately from original benchmark reproduction. These approaches provide the reconstruction foundations used here.

Uncertain-input GP inference already accounts for location uncertainty through linearization, latent inputs and posterior integration (McHutchon & Rasmussen, 2011; Damianou et al., 2016; Jadaliha et al., 2013). Our moment-based comparison places an alternative treatment of location uncertainty inside the shared reconstruction solver. We study auxiliary likelihoods for missing-component, recurrent reconstruction and their use in selection versus correction.

Excluding observed components has established foundations in self-supervised denoising (Batson & Royer, 2019). Here sensor-block exclusion applies to the auxiliary location fit; the shared reference and history retain sensor dependence. FreqNO-DPS calibrates neural-operator residuals for diffusion guidance with known sensor sites and complete component histories (Perrone et al., 2026). We instead calibrate visible residuals for latent-location weights in a causal kernel-correction program.

Loss-calibrated and semi-modular inference distinguish estimation from decision quality and control module feedback (Lacoste-Julien et al., 2011; Carmona & Nicholls, 2020). Taylor GP updates following improved location estimates also have precedent (Qureshi et al., 2026). Our analysis tracks probability-weight changes through the implemented covariance, solve, closure and cap, and pairs that response with controlled stage interventions. This connects the location estimator to the behavior of the complete reconstruction program.

3 Method

Figure 1 summarizes one step of AB. The corrected field feeds the next prediction; calibration parameters remain fixed throughout the rollout.

3.1 Observations and References

At step h, sensor m supplies visible components $M_m y_m$, a reported location, and a prior π_{mj} over neighboring sites x_{mj}. Inference has two initial complete fields and previous reconstructed outputs.

The current dense field and actual sensor positions are reserved for evaluation. The candidate support is supplied with the observations; multiplying by the prior cannot recover a site assigned zero prior probability.

The pretrained neural predictor and coefficient history produce reference candidates indexed by $\alpha \in \{0, 1/2, 1\}$. Each is based on a mixture of the initial field and the closed neural prediction, with its own causal coefficient state. Reference and solver configurations are selected by two repeated sensor bipartitions, swapping proposal and held halves to form four panels. With 64 sensors, each proposal uses 32 sensors. Held mixture deviance selects the solver configuration within a reference; a common scaled held score compares references. The final fit uses all 64 sensors.

3.2 Two location-weight generators

Generator A forms an eight-dimensional regularized auxiliary correction around the current reference, with the frozen ridge multiplier 0.1. On cylinder, its directions are the first eight closed global modes of the training-derived bank; Appendix A.2 records basis construction and fixed design choices. A Gaussian working likelihood over the supplied neighborhood gives weights proportional to

$$q^A_{mj} \;\propto\; \pi_{mj} \exp\left[-\frac{\|M_m(y_m - \widetilde{r}_A(x_{mj}))\|^2}{2v} \right], \tag{1}$$

where v is the visible residual variance around the reference. The auxiliary increment is used to interpret locations rather than added separately to the final field.

Generator B fits auxiliary estimates with one of four fixed sensor blocks excluded, using the estimate excluding sensor m to evaluate its location likelihood. This exclusion applies to the auxiliary fit; the common predictive reference and history can still depend on the sensor. Residual means and variances, conditioned on the component mask, are calibrated on training data and frozen before evaluation. Writing $\widetilde{r}_{-b(m)}$ for the auxiliary field fitted without sensor m's block, set $e_{mjc} = (y_{mc} - \widetilde{r}_{-b(m),c}(x_{mj}))/\sqrt{v}$. The weights satisfy

$$q^B_{mj} \;\propto\; \pi_{mj} \exp\left[-\frac{1}{2} \sum_{c \in M_m} \left\{ \frac{(e_{mjc} - \mu_{M_m,c})^2}{s^2_{M_m,c}} + \log s^2_{M_m,c} \right\} \right]. \tag{2}$$

This construction changes the candidate weights; it leaves the scoring equations unchanged. Appendix A.3 specifies which quantities are shared and which are recomputed in each auxiliary fold.

3.3 Common full kernel correction

Let H_q average visible field components using q, U be the selected background basis, and K the Gaussian kernel. Write $X_q = H_q U$, $K_q = H_q K H_q^*$ and $e = y - H_q r$. For a working observation covariance C_q and positive penalties λ, κ, the implemented solve is

$$A_q = K_q + \kappa C_q, \qquad J_q = X_q^\top A_q^{-1} X_q, \tag{3}$$

$$z = (J_q + \lambda/\kappa\, I)^{-1} X_q^\top A_q^{-1} e, \tag{4}$$

$$\beta = A_q^{-1}(e - X_q z), \qquad f = r + Uz + K H_q^* \beta. \tag{5}$$

The field then undergoes the shared discrete closure and increment cap. Thus changing q changes the sampling operator, Gram matrix, cross-kernel and working covariance consistently.

The common grid has 109 settings: a no-correction candidate and combinations of background rank $8, 16, 32$, six background ridge multipliers, three kernel regularization scales and two kernel lengths. The kernel term extends the correction beyond the background basis.

3.4 Stage assignment and recurrence

We denote selection and final correction generators by two letters. AA uses A in both stages; BB uses B in both. Our AB procedure selects the reference and configuration using A on its own current history, then recomputes only the final full-observation correction using B. BA is the reverse

intervention. After producing its output, each method updates the registered coefficient states and feeds its own completed field to the next neural prediction.

The GP moment comparator uses the same priors and candidate opportunities, with correct second moments for the single latent location shared by a sensor's visible components. Accurate-location conditions provide a common-solver numerical check.

4 From Location Probabilities to Field Responses

We analyze the implemented inverse map at a fixed reference and configuration. We track a change of location weights through the solve, physical closure and amplitude cap to the final field.

4.1 The implemented correction and its final loss

Hold the reference r, candidate configuration, observation mask and prior-normalized background penalty fixed. Let U be the background basis and $G = K + (\kappa/\lambda)UU^\top$, with $\lambda, \kappa > 0$. For location weights q, write H_q for the weighted observation operator and C_q for the implemented working covariance. The raw production solve is

$$S_q = H_q G H_q^* + \kappa C_q, \qquad D_q = G H_q^*, \qquad e_q = y - H_q r, \qquad f_q = r + D_q S_q^{-1} e_q. \tag{6}$$

The working covariance C_q combines location-weighted reference variability with the common residual normalization. The no-correction candidate bypasses this solve and returns r.

For two weight choices q_A, q_B, denote differences by Δ and let $w_A = S_A^{-1} e_A$. With $M_q = H_q G H_q^*$, the following four terms sum to the raw field difference:

$$\begin{aligned} d_{\text{prop}} &= (D_B - D_A) w_A, & d_{\text{res}} &= D_B S_B^{-1} \Delta e, \\ d_{\text{Gram}} &= -D_B S_B^{-1} \Delta M \, w_A, & d_{\text{cov}} &= -D_B S_B^{-1} \kappa \Delta C \, w_A, \\ f_B - f_A &= d_{\text{prop}} + d_{\text{res}} + d_{\text{Gram}} + d_{\text{cov}}. \end{aligned} \tag{7}$$

This exact finite-change identity follows from $S_B(w_B - w_A) = \Delta e - \Delta S \, w_A$. It separates changes in field propagation, observation residuals, the Gram matrix and the working covariance without requiring a small perturbation. The decomposition uses A as its anchor; its individual term sizes are not unique causal effects.

The implemented output also includes the common affine closure $c_q = u_0 + P(f_q - u_0)$ and the amplitude cap

$$F_q = r + s_q(c_q - r), \qquad s_q = \min\{1, \tau_r / \operatorname{RMS}(c_q - r)\}, \tag{8}$$

where $s_q = 1$ at zero update and τ_r is the common reference-based budget. Thus the final difference is

$$d = F_B - F_A = s_B P \sum_j d_j + (s_B - s_A)(c_A - r). \tag{9}$$

Closure changes the error geometry from Ω to $P^* \Omega P$, with an affine target offset (Appendix X.1). The cap additionally changes relative scales; its full-grid RMS norm differs from the fluid-domain evaluation norm.

For a target u and a fixed evaluation weight Ω, the final quadratic loss difference is

$$\|F_B - u\|_\Omega^2 - \|F_A - u\|_\Omega^2 = 2\langle F_A - u, d\rangle_\Omega + \|d\|_\Omega^2. \tag{10}$$

B improves precisely when its final difference is sufficiently aligned against A's remaining error. Accurate prediction of d alone does not determine that alignment. For diagnostics, writing Eq. (9) as $d = \sum_j t_j$ gives the exact allocation $\sum_j \langle t_j, 2(F_A - u) + d\rangle_\Omega$. This uses the evaluation target and is not an observation-only decision rule. These identities apply directly to the implemented solve. A residual-based bound controls response-computation error after closure and cap (Appendix D); Section 5.7 evaluates it separately from field benefit. The conditional-location approximation in Appendix I.1 is a supplementary diagnostic, not the AB solver.

Table 1: Relative NRMSE reduction of AB over independently calibrated reference alternatives. Positive favors AB. Each cell reports FNO / LocalNO; the later window changes time and observation seed on the same three cylinder cases (eight conditions, H1–20). Both alternatives share the reconstruction solver and use independently fitted reference calibration, frozen across windows.

Matched alternative	Original window	Later window
A–R: A selects, reference weights fit	13.27% / 11.19%	10.18% / 7.22%
R–R: reference weights select and fit	8.34% / 9.03%	4.12% / 4.40%

5 EXPERIMENTS

5.1 EVALUATION PROTOCOL

The main panel contains cases 0019, 0043 and 0048, FNO and LocalNO reference models and observation-seed offset 5100. Each task covers eight conditions and twenty recurrent steps, denoted H1–20. These cases have prior development exposure. The primary experiments use 64 sensors and independent method histories. Cross-physics evaluations include five reserved natural RBC trajectories, shear flow, PDEArena, liquid-metal magnetic fields and two three-dimensional JHTDB panels. Their complete comparisons and protocols are retained in Appendices N, Z and E.

Field NRMSE is $\|\widehat{u} - u\|_2 / \max(\|u\|_2, 10^{-12})$ over all components on the evaluation domain. We average over conditions and horizons within each physical unit, then weight units equally. Backbones are reported separately; observation repetitions are not additional physical samples. References, calibration parameters and candidate grids remain fixed. The cylinder panel supports the component and public-architecture comparisons; the cross-physics panels examine transfer and failure conditions.

Comparison roles. Adaptations of published architectures retain their own inference and recurrent feedback. Matched alternatives share the kernel program to test location inference; AA/AB/BB are internal variants. Only the latter two groups share the 109-setting selector. Table 9 records source, adaptation, training/calibration and evaluation scope for every comparison group. We report accurate-location, full-component observations as the external anchor and list location-jitter and missing-component extensions separately.

5.2 MATCHED ALTERNATIVES WITHIN THE RECONSTRUCTION PROGRAM

To test whether calibration of the predictive reference suffices, we fit reference-residual calibration independently, with the same training split, mask conditioning, estimator family and sample opportunities as auxiliary calibration. Both alternatives use the common complete kernel solver. A–R retains A selection and replaces only final weights by calibrated reference weights; R–R uses calibrated reference weights for both selection and final fitting. Each alternative updates its own history and has the same 109 settings and three reference opportunities as AB.

Case 0048 favors AB over R–R by 12.32–14.88% across both windows and backbones. Case 0043 instead favors R–R in the later window by 4.06%/6.84%; case 0019 is mixed. Allowing reference weights to participate in selection narrows the gap, through selection and recurrent feedback together. All case results are retained in Appendix C.

A second matched alternative uses GP prior moments. It incorporates the prescribed location distribution in its first and second moments, including the location shared across a sensor's visible components. It shares references, kernel/background structure, candidate opportunities, scoring and postprocessing, but not the auxiliary weight generator. AB's aggregate advantage over this comparator is 2.28%/1.22% in the original cylinder window and 0.23%/0.38% later. In case 0048, GP moments remain better in both windows, despite AB's clear advantage over R–R. Thus the benefit over calibrated reference weights does not imply the same ranking against GP moments.

Table 2: Frozen AB on the three cylinder cases, H1–20. Relative NRMSE reductions are computed from panel means. AA and BB are internal variants; GP moments shares the reconstruction program.

Backbone	AB NRMSE	vs. AA	vs. BB	vs. GP moments
FNO	0.021115	5.25%	1.51%	2.28%
LocalNO	0.020513	3.40%	0.71%	1.22%

Table 3: Component reductions over AA (%), averaged over the same three cylinder cases and H1–20. Each arm selects with A and uses its own recurrent history. Only the observation seed changes between batches. Positive values indicate lower error.

Observation batch	Backbone	Cross-fit only	Calibration only	Both (AB)
Original	FNO	1.23	4.07	5.25
Original	LocalNO	3.19	2.99	3.40
Repeat	FNO	-0.93	2.12	2.29
Repeat	LocalNO	2.31	3.37	1.63

5.3 Repeated gains on the cylinder panel

AB improves over AA in all three cylinder cases for both backbones, and over BB and GP moments in two cases. The repeated component experiments below distinguish this aggregate result from the effect of each design choice. A same-checkpoint replay gives closed-reference errors of 0.14897/0.14504 for FNO/LocalNO, versus 0.02112/0.02051 for AB. This measures the complete reconstruction system; the location-estimator contribution requires the matched alternatives and ablations (Appendix N.1).

5.4 Component and stage ablations

Table 3 compares the two component changes across observation batches. Calibration improves all six case–backbone means in both batches. Cross-fitting improves six in the original batch and four in the repeat; the combination has the lowest aggregate error for both backbones originally, but only for FNO in the repeat. Relative to independently calibrated reference weights, AB retains repeated gains of 8.30%/3.34%, concentrated in 0048. Appendices F, F.3 and F.4 give case effects and leave-one-case sensitivity.

Stage assignment is less stable. The original window favors AB over BB by 1.51%/0.71%; a later nonoverlapping window favors BB by 1.54%/0.70%, with BB winning five of six groups. AB still improves over AA by 1.79%/0.70% there. Thus the modified location estimator has value in both assignments, while the preference for A selection does not persist. The same cases are used in both windows (Appendix P).

A post-hoc full-recurrence control sets residual means to zero and variances to 0.25. Full-fit sharpening recovers the learned-calibration gains; AB improves over this simpler arm by 0.15% with FNO but loses 2.28% with LocalNO. Within cross-fitting, learned statistics improve the means by 2.83%/0.85%, winning three of six groups (Appendix A.6). Thus the likelihood adjustment has repeatable value, while the complete combination is not uniformly needed.

5.5 Complete systems: adaptations of published architectures

We compare AB with task adaptations of the authors' implementations of RecFNO, FLRONet, Senseiver and PalSB (Zhao et al., 2024; Dang & Nguyen, 2025; Santos et al., 2023; Li et al., 2025). All receive causal histories and visible sparse observations, retain their own recurrent outputs and select checkpoints on separate calibration cases. Interfaces, budgets and matched-observation checks are specified in Appendix B.

Table 10 reports the original-budget RecFNO and FLRONet comparisons. AB has lower error on all three clean cases against all four adaptations. The original-budget Senseiver and PalSB clean means are 0.17275 and 0.30110, versus 0.01316/0.01234 for AB/FNO and AB/LocalNO (Appendix B.4).

Table 4: Direct location diagnostics on jittered observations. Lower NLL, Brier and field NMSE are better; higher MAP accuracy is better. Cylinder uses cases 0019/0048 (32 queries); RBC uses trajectories 5/6 (16 queries). Both include two backbones and H1/H10. Means weight physical units equally. Appendix A.2 defines the scores and sampling.

Panel	Weights	NLL	Brier	MAP (%)	Field NMSE
Cylinder	A	1.8936	0.8138	44.38	0.001175
Cylinder	B	1.8044	0.7774	43.12	0.001141
RBC	A	1.9421	0.8153	28.22	0.141250
RBC	B	1.9449	0.8119	28.42	0.141470

Bounded RecFNO and FLRONet continuations improve their selected models; the two Senseiver continuations are reported separately in Table 16; PalSB's rollout-based physical finetuning leaves CAL selecting the pre-finetuning checkpoints. A second RecFNO seed retains the clean-system advantage. Under uncertain locations, AB additionally uses the supplied position law; clean removes this information difference. These are complete-system comparisons under the reported budgets. Tables 1 and 3 isolate the location estimator.

5.6 Location inference at a fixed solver

We evaluate the location probabilities directly on a separate observation batch, holding AA's reference, history and solver fixed (Table 4). On cylinder, B improves NLL and Brier score while reducing MAP accuracy: probability quality and point-location accuracy give different answers. RBC has a small Brier improvement but worse NLL and field error. These diagnostics complement the complete recurrent comparisons.

Across these additional observations, B improves both location scores in 28/32 cylinder queries and 8/16 RBC queries. Both scores improve yet final field NMSE increases in 4/32 and 1/16 queries, respectively. Among those exceptions, mean excess NMSE is 2.39×10^{-4} on cylinder and 1.62×10^{-4} on RBC. These frequencies describe the two physical units in each panel; Figure 2 illustrates one separate archived path.

5.7 Mechanism interventions and failure paths

Separating final fitting from selection. At a fixed donor history and reference, let $F_X(j)$ be the final postprocessed field using weights $X \in \{A, B\}$ and configuration j, and let j_X be the configuration selected using X. For a common field loss L, the conditional change satisfies

$$L(F_B(j_B)) - L(F_A(j_A)) = \underbrace{L(F_B(j_A)) - L(F_A(j_A))}_{\text{final-weight effect}} + \underbrace{L(F_B(j_B)) - L(F_B(j_A))}_{\text{selection effect at final B}}. \tag{11}$$

The first term changes only final weights; the second changes the selected configuration while retaining B for the final fit. Equations (7)–(10) describe the field response underlying the first term. Held-sensor scoring can favor a configuration whose full-field loss is higher, offsetting a beneficial weight change.

Controlled evidence. At H1 and H10 of AA trajectories, all 176 conditional AA/AB/BA/BB queries pass replay checks. Changing only final weights improves mean cylinder NRMSE at H1; then allowing configuration reselection reduces that benefit. At H10 the reselection effect differs across backbones, and RBC's fixed-configuration mean changes from a benefit at H1 to a small loss. These are algorithmic interventions at a fixed reference and history. Of 37 configuration switches, only one harms final A but helps final B in the same query; larger panel-average reversals primarily reflect heterogeneous effect sizes. Appendix G reports both paths of Eq. (11).

Postprocessing and recurrence. A production replay reproduces the final-field differences and shows that closure or capping can reverse A/B rankings (Appendix X). Independent recurrence adds changing references and feedback: AB beats BB in the original cylinder aggregate, then loses in the later window. The three experiments therefore address complementary questions: how weights

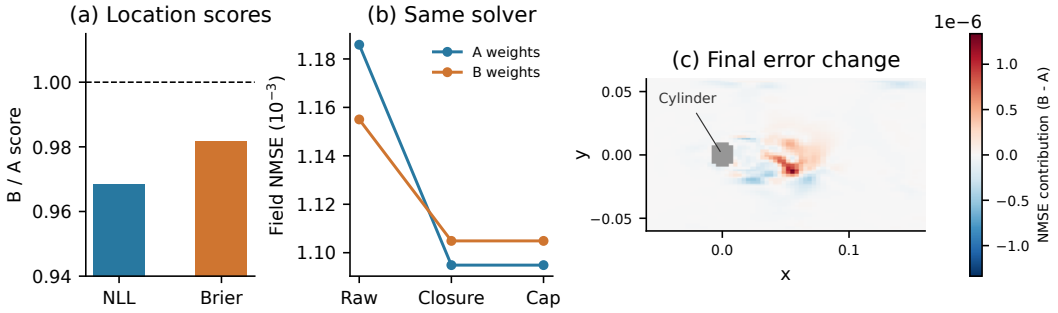

Figure 2: A same-query response: case 0043/FNO, H10, jitter with missing v. Reference, history and configuration are fixed. B improves both location scores and raw field NMSE, but closure reverses the field-error ranking; the cap is inactive. Panel (c) shows each grid point's contribution to the final NMSE difference (B minus A): red indicates harm and blue improvement. Gray marks the excluded cylinder; axes preserve the physical aspect ratio.

change a fixed fit, how selection alters that fit, and whether the complete algorithm retains the gain. They use separate panels; Appendix G.2 specifies the controlled quantities and inference supported by each.

Figure 2 isolates one such path. NLL decreases by 0.0634 and Brier by 0.0155. The raw field NMSE decreases by 3.09×10^{-5}, whereas the closed update increases it by 1.00×10^{-5} relative to A. Closure reduces both methods' absolute errors, but benefits A more. Of the final change, $+1.18 \times 10^{-5}$ lies in missing v and -1.80×10^{-6} in observed u. Thus the projection of a changed update can reverse its relative benefit without any reselection or feedback.

5.8 Failure conditions

Panel averages hide substantial losses. Case 0048/LocalNO with jitter and a missing component has 7.80% higher NRMSE under AB than AA and 33.27% higher than GP moments. Shear also contains condition-level harm despite every task mean improving over GP. RBC favors AA over AB on every transfer trajectory, although AB still improves on the same-checkpoint predictive reference. Shear and liquid-metal panels favor AB over the shared GP moments on average, while PDEArena adds little over AA. With stronger PDEArena references, the original correction can also worsen the reference itself; a calibration-only amplitude adjustment reduces but does not eliminate this harm (Appendix N.1). Appendix R reports the complete failure examples and Table 26 the registered 5% harm counts.

5.9 Where accepted corrections fail

To locate acceptance errors, we replay 40 parent queries on exposed MHD and isotropic objects with both backbones, retaining the selected configurations and donor histories. Among the original score's 20 harmful decisions, all four half-observation candidates are already worse than their respective references in field risk. The final refit therefore does not create the harmful sign in this diagnostic panel.

Two distinct mismatches occur. In 12 decisions, the location-mixture score favors correction even though held-sensor error at the true positions increases. In the other eight, true-position held-sensor error decreases while full-domain error on the visible components increases. Both distinctions use target-assisted diagnostics and are not deployment tests. The latter also occurs with accurate positions, so improving location inference alone cannot eliminate acceptance errors. Appendix J gives the replay design and an additional spatial-sampling check.

Checking response calculations. Residual-certified inverse solves meet the postprocessed numerical precision criteria on all 32 resolvable validation responses, versus 24 for a second-order approximation. Appendix K gives the bound checks, unresolved cases and computational costs.

Three-dimensional scope. The current method was also evaluated on six JHTDB MHD spatial objects and two isotropic objects. The three uncertain-location condition means favor AB over AA, BB and common GP on MHD; isotropic favors AA and GP. Fixed-history replays show that isotropic harm already occurs in final weight replacement and is accompanied by worse location likelihood, before changing selection or recurrent history. These are object-specific observations rather than a rule based on physical-field labels. Appendix E gives all results, repeated-window dependencies and the additional NIGP-inspired comparison.

The measured H1 implementation overhead of AB over AA is 8.9% on cylinder and 12.4% on RBC. Appendix Q reports the matched timing protocol and memory use. These costs are distinct from response-certificate computation.

6 Discussion and Limitations

The cylinder results support adjusting auxiliary likelihoods. AB is the frozen version selected on earlier development data. On a new task, compare full-fit sharpening, calibration, cross-fitting and stage assignment using complete training/validation rollouts and the intended field metric. This selection requires validation fields. The auxiliary-field gain over calibrated reference weights is concentrated in 0048.

Closure evaluates a location-induced update in a transformed error geometry (Appendix X.1); the paired intervention locates the resulting harm in the missing component. Acceptance replay separates location-score mismatch from sparse-to-field mismatch, which persists at true locations. These distinctions apply beyond this estimator; their frequency and sign depend on the solver, closure and sampling. Numerical response accuracy addresses a separate question.

Reference quality changes the value of correction. AB improves the RBC reference but loses to AA. Strengthening the reference preserves correction gains on liquid metal but can produce harmful corrections on PDEArena. These reference-replacement interventions retain the original downstream calibration and state process.

The closest reference and public-architecture comparisons cover three cylinder cases under finite training budgets, with two RecFNO seeds and one FLRONet seed. Several panels have development exposure, and repeated observation batches reuse physical cases. The evaluated location supports contain the actual sensor sites; performance outside those supports remains untested.

7 Conclusion

Calibrated auxiliary location inference gives repeated gains on the cylinder panel, with component and matched-reference comparisons identifying their sources. The production-response analysis and controlled interventions explain how changes in location probabilities propagate through the implemented reconstruction and where their benefit can be lost. Evaluating the final correction and its selector separately reveals the roles of physical closure, sparse scoring and stage assignment. Together, these results establish conditional improvements and a concrete account of their behavior in recurrent reconstruction.

Reproducibility Statement

Appendices A.3, Y and Z specify algorithms, equations and data splits; Appendix B documents comparators. Appendix A describes matched-observation audits, a fixed-query CPU replay and the verified four-arm H1–20 recurrence package.

AI Use Statement

Generative AI tools assisted with research hypotheses, experimental design, mathematical derivations and proof writing, implementation, analysis, literature search, figures, and writing. Results were checked against recorded experiments and artifacts. The authors take responsibility for the final submission.

REFERENCES

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

Christian Carmona and Geoff Nicholls. Semi-modular inference: enhanced learning in multi-modular models by tempering the influence of components. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4226–4235, 2020. URL https://proceedings.mlr.press/v108/carmona20a.html.

Peng Chen, Umberto Villa, and Omar Ghattas. Taylor approximation and variance reduction for pde-constrained optimal control under uncertainty, 2018. URL https://arxiv.org/abs/1804.04301.

Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17 (42):1–62, 2016. URL https://jmlr.org/papers/v17/15-293.html.

Hiep Vo Dang and Phong C. H. Nguyen. Deep operator learning for high-fidelity fluid flow field reconstruction from sparse sensor measurements. *Journal of Computing and Information Science in Engineering*, 26(1):011007, 2025. doi: 10.1115/1.4070332.

Eirik Endeve, Doug Stefanski, Marc-Olivier G. Delchini, Stuart Slattery, Cory D. Hauck, Bruno Turcksin, and Sergey Smolentsev. A full-induction magnetohydrodynamics solver for liquid metal fusion blankets in Vertex-CFD. *arXiv preprint arXiv:2511.15549*, 2025. URL https://arxiv.org/abs/2511.15549.

R. Everson and L. Sirovich. Karhunen–loeve procedure for gappy data. *Journal of the Optical Society of America A*, 12(8):1657–1664, 1995. doi: 10.1364/JOSAA.12.001657.

Helin Gong, Yvon Maday, Olga Mula, and Tommaso Taddei. PBDW method for state estimation: error analysis for noisy data and nonlinear formulation. *arXiv preprint arXiv:1906.00810*, 2019.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

Mahdi Jadaliha, Yunfei Xu, Jongeun Choi, Nicholas S. Johnson, and Weiming Li. Gaussian process regression for sensor networks under localization uncertainty. *IEEE Transactions on Signal Processing*, 61(2):223–237, 2013. doi: 10.1109/TSP.2012.2223695. URL https://pubs.usgs.gov/publication/70043209.

Simon Lacoste-Julien, Ferenc Huszár, and Zoubin Ghahramani. Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 416–424, 2011. URL https://proceedings.mlr.press/v15/lacoste_julien11a.html.

Zeyu Li, Hongkun Dou, Shen Fang, Wang Han, Yue Deng, and Lijun Yang. Physics-aligned field reconstruction with diffusion bridge. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=D042vFwJAM.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Miguel Liu-Schiaffini, Julius Berner, Boris Bonev, Thorsten Kurth, Kamyar Azizzadenesheli, and Anima Anandkumar. Neural operators with localized integral and differential kernels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32576–32594, 2024. URL https://proceedings.mlr.press/v235/liu-schiaffini24a.html.

Yvon Maday and Tommaso Taddei. Adaptive PBDW approach to state estimation: Noisy observations; user-defined update spaces. *SIAM Journal on Scientific Computing*, 41(4):B669–B693, 2019. doi: 10.1137/18M116544X.

Yvon Maday, Anthony T. Patera, James D. Penn, and Masayuki Yano. A parameterized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics. *International Journal for Numerical Methods in Engineering*, 102:933–965, 2015. doi: 10.1002/nme.4747.

Andrew McHutchon and Carl Edward Rasmussen. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://papers.neurips.cc/paper_files/paper/2011/hash/a8e864d04c95572d1aece099af852d0a-Abstract.html.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Regaldo-Saint Blancard, Francois Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-1430.

Niccolò Perrone, Fanny Lehmann, Stefania Fresca, and Filippo Gatti. Correcting neural operator spectral bias via diffusion posterior sampling with sparse observations, 2026. URL https://arxiv.org/abs/2606.03936. Preprint.

Muzaffar Qureshi, Tochukwu Elijah Ogri, Kyle Volle, and Rushikesh Kamalapurkar. A taylor series approach to correct localization errors in robotic field mapping using gaussian processes, 2026. URL https://arxiv.org/abs/2601.20149. Preprint.

Javier E. Santos, Zachary R. Fox, Arvind Mohan, Daniel O'Malley, Hari Viswanathan, and Nicholas Lubbers. Development of the senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence*, 5:1317–1325, 2023. doi: 10.1038/s42256-023-00746-x.

Doug Stefanski, Paul Laiu, Matthew Norman, John Gounley, and Pei Zhang. Liquid-metal magnetohydrodynamic lid-driven cavity flow dataset. Zenodo, version 1, 2026. URL https://zenodo.org/records/19857524.

Xiaoyu Zhao, Xiaoqian Chen, Zhiqiang Gong, Weien Zhou, Wen Yao, and Yunyang Zhang. RecFNO: A resolution-invariant flow and heat field reconstruction method from sparse observations via fourier neural operator. *International Journal of Thermal Sciences*, 195:108619, 2024. doi: 10.1016/j.ijthermalsci.2023.108619.

A CLAIM AND EVIDENCE MAP

The main evidence follows three questions: what changes the algorithm, what changes its error, and what repeats in complete rollouts. Algorithm definitions and calibration are in Appendices A.3 and Y; component and simple-scale comparisons are in Appendices F and A.6. For the mechanism chain, read Appendix G.2, then the production replay in Appendix X and acceptance audit in Appendix J. External comparisons and 3D transfer are in Appendices B and E. The conditional-mixture approximation and directional precision studies are supplementary numerical diagnostics, not additional AB modules.

Table 5: Which comparison supports which claim. Conditional interventions and independent recurrence answer different questions.

Claim	Direct comparison	Supported scope
The complete system improves over tested public architectures	Current AB versus the pinned author-core RecFNO and FLRONet task adaptations, with paired observations and native recurrence	All three clean case means favor AB against each adapter. RecFNO has two training seeds and FLRONet one, with finite budgets; it does not isolate an auxiliary effect.
Auxiliary location inference can add to reference calibration	Independently calibrated A–R and R–R alternatives, complete recurrence in two windows	Aggregate gains in both windows and an additional observation seed, concentrated in 0048; mixed 0019/0043 effects. GP moments still outperform AB on 0048.
Likelihood adjustment improves AA	AA, cross-fit only, calibration only, AB and fixed-scale controls, with independent histories	Calibration helps all six groups in both batches. Fixed sharpening recovers the full-fit calibration gains in the repeat. Learned statistics improve the cross-fitted aggregate versus fixed scale, winning three of six groups. Mask grouping and the necessity of complete AB remain unestablished.
Stage assignment changes reconstruction	Conditional AA/AB/BA/BB and independent AB versus BB; both AB and BB use the modified estimator	Later-window BB wins five of six groups, limiting a fixed AB-assignment claim. This internal comparison does not negate the separate cross-fitting and calibration ablations.
The complete procedure competes with GP moments	AA, AB, BB and GP under matched candidate and observation opportunities	AB improves original-batch panel means on cylinder, shear, liquid-metal magnetic fields and 3D MHD; RBC, PDEArena and 3D isotropic turbulence retain negative comparisons. The additional cylinder seed favors GP with LocalNO. This is not a comparison of all joint GP methods.
Production response precision can be controlled	Frozen residual-based stopping and matched direct solves on six tasks	32/32 resolvable queries pass; second order passes 24/32. A same-query guard repair excludes unresolved responses. This is numerical precision, not target-field benefit.
Sparse score gains can miss field harm	Exact candidate replay with true-position held-score and dense-field diagnostics	Among 20 harmful decisions in 40 MHD/isotropic parent queries, 12 show location-score mismatch and eight show true-position sparse-to-field mismatch. All half-fit candidates are already harmful; these are target-assisted diagnoses.

A.1 Evidence Organization and Reproduction

Comparisons within each task use the same generated observations. Stored checkpoint and calibration hashes identify the fitted estimators; observation fingerprints pair the outputs; per-task receipts and table-export checks connect the reported means to those outputs. The public-architecture evaluation has 480 queries per training seed, while the independently calibrated reference comparison covers 24 tasks across two windows. Component and observation repeats are reported with their physical cases rather than counted as new datasets. The three-dimensional tables likewise distinguish spatial objects from overlapping windows and repeated model tasks.

Implementation equations and data splits appear in Appendices Y and Z; comparator sources and interfaces are in Appendix B. Appendix A.3 gives the recurrent and auxiliary-fold algorithms. A separate CPU replay package covers one cylinder query and one RBC query: it refits the full and held-block auxiliary fields, recomputes A/B weights, solves the cached production systems, applies the stored closure action and actual cap, and checks the resulting fields against the archive. It reproduces these fixed-query calculations, rather than training, reference selection or independent recurrence. The manuscript-source archive contains LaTeX and figure assets.

A complete pretrained recurrence package covers case0048/FNO and the four component arms: AA, cross-fitting only, calibration only and AB. Each arm recomputes all eight conditions over H1–20 on its own history, including neural reference branches, candidate selection, the coefficient state, correction and postprocessing. In an independent directory, with access to original project assets blocked, all 640 results reproduce the archived reference/configuration choices, observation fingerprints and NRMSEs. A separate check regenerates all 160 sensor inputs and supplied priors exactly in the recorded environment. The bundle includes the frozen checkpoint, data window, calibration and state assets, a consolidated case configuration, environment versions and a hash-checked entry point. Each arm took about 18 minutes in the tested CUDA environment. This is complete inference from pretrained assets; fresh model training and all-dataset reproduction are outside this bundle.

Artifact entry points and coverage. The two experiment archives are separate from the manuscript-source archive. Their exact filenames are `ICLR_RECURRENT_CASE0048_20260926.zip` and `LOCATION_RESPONSE_CPU_REPLAY_20260926.zip`. Each has a root `README.md`. After extraction, the recurrence archive uses Python 3.10, CUDA and the package versions in `environment.json`; run `python reproduce.py --arm combined` for AB. Replace `combined` by `aux`, `crossfit`, or `calibrated` for the other component arms, and use `python check_observations.py` to check the generated observation packet. Set OMP, MKL and OpenBLAS thread counts to one as specified in the README. The CPU archive uses `requirements.txt`; run `python replay.py`.

The recurrence outputs regenerate the case0048/FNO contributions to the original component comparison, including its per-condition H1–20 fields, choices and errors. They do not regenerate the other case/backbone means, repeated observation batches, fixed-scale control, external-model training or cross-physics tables. The CPU fixtures verify two fixed-query calculations rather than a complete table. `CASE_CONFIG.json`, the manifests and entry-point checks identify the bundled inputs. An index accompanying the archives separates these executable checks from the manuscript's full experimental record.

A.2 Location Diagnostics and Parameter Provenance

For sensors with more than one admissible site, location NLL is the mean of $-\log q_{m,j_m^*}$, Brier score is the mean of $\sum_j (q_{mj} - \mathbf{1}[j = j_m^*])^2$, and MAP accuracy is the fraction whose maximum-weight site equals the actual site. The actual sites are used only for evaluation. Field NMSE is squared NRMSE. We average within each physical unit before averaging units. These are scores of the working location probabilities, not a test of exact Bayesian posterior calibration.

The archived study has 69 selected-query replays and 96 queries from a fresh observation batch. Table 4 uses the 48 uncertain-location queries from that fresh batch (observation-seed offset 6100); accurate-location controls are excluded. The reference and history come from AA at H1/H10. Cylin-

der uses cases 0019/0048; RBC uses reserved natural trajectories 5/6. Of these queries, 4/32 on cylinder and 1/16 on RBC improve both NLL and Brier under B yet increase final field NMSE. Queries share physical units and recurrent histories; these fractions describe this panel.

For context, uncalibrated reference weights on cylinder give NLL 1.9160, Brier 0.8170, MAP accuracy 45.41%, and field NMSE 0.001498. The RBC values are 1.9350, 0.8138, 31.54%, and 0.140430, respectively. These weights are distinct from the independently calibrated R–R comparator. The same study tested a directional prediction that removing a component and re-estimating weights would increase B's excess risk over A relative to freezing full-component weights. The mean contrasts were negative on both panels, so this prediction was not supported. True-position weights reduced mean final NMSE in the missing-component diagnostic, by 0.000204 on cylinder and 0.003908 on RBC.

Figure 2 is a post-hoc illustration from the replay set, selected to show a raw-to-closed ranking reversal. Its stored field arrays reproduce the reported risks and archived solver outputs. The figure and fresh-batch frequencies serve different purposes: one isolates an observed path, the other checks whether location-score improvements can coexist with field harm on additional observations.

Basis construction. On cylinder, the 32-direction background bank concatenates the first 16 training-derived global increment modes with 16 geometry-only local potential fields. Eight fixed spatial centers are used at each of two scales, followed by the dataset's discrete closure. The auxiliary estimator uses the first eight closed global directions. Its basis is fixed before inference; it is not learned from evaluation residuals. Other panels retain their registered training-derived banks; the PDEArena and liquid-metal constructions are specified in Appendix Z. The background bank and auxiliary bank are shared by A and B.

Fixed design choices. The auxiliary rank eight and ridge multiplier 0.1 were inherited from development of the original estimator. Four index-modulo sensor blocks and the calibrated variance interval $[0.25, 16]$ were fixed when B was introduced; there was no variance-clip search. Residual means and variances are fitted on the registered training cases, separately for full and cross-fitted auxiliary estimates, and then frozen. The 109-setting grid and 0.25 increment cap are inherited common-solver settings. These evaluations measure the frozen choices, rather than optimality of each constant. Appendix Y gives their equations and normalization.

Matching backbone results. The $\alpha = 0$ reference omits the neural prediction but retains the initial field and causal coefficient-state correction. When both backbone runs select this branch and the same solver, their reconstruction errors can coincide. This explains why identical entries need not indicate duplicated neural predictions. On case 0019's clean H1–20 comparison, both backbones select $\alpha = 0$ throughout and have matching AB configurations and errors. MHD has extensive use of this branch as detailed in Appendix E; those repeated outputs do not provide independent evidence about neural-backbone quality.

A.3 One Recurrent Step and the Auxiliary Folds

Frozen inputs. The neural predictor, background bank, coefficient-state model, calibration tables, reference choices, candidate grid and sensor-partition rule are fixed before evaluation. Calibration uses true sites of training observations. A reconstruction step receives visible values, masks, reported sites, supplied priors, the initial complete field and the method's own previous outputs and coefficient states.

AB step.

1. Form neural predictions and proposal predictions from the causal inputs. For each α, construct its reference and proposal fields using that branch's pre-step coefficient state.
2. For each reference and each of the four proposal/held panels, use A on the proposal sensors to generate candidate fields. Evaluate their held location-mixture deviance. Select the solver within each reference and compare references using the common scaled score.

3. Retain the selected reference and configuration. On the full sensor panel, compute B's location weights by the fold procedure below. Apply the common kernel solve with these weights in H_q, K_q, the cross-kernel and C_q.
4. For a nonzero correction candidate, apply closure relative to the initial field and the reference-based increment cap; the no-correction candidate returns its reference directly. The resulting complete field becomes the next neural input.
5. Assimilate the current observations into each registered α coefficient state. These states are carried to the next step; they are not updated separately inside auxiliary folds.

AA replaces B by A in step 3. BB also uses B in step 2 and runs on its own history. Neither comparison reuses AB's future outputs.

B on a fitting panel I. Compute the reference residual scale v_I from all visible values in I, as in Appendix Y. Partition its ordered sensor indices modulo four. For each block b, fit the rank-eight, ridge-0.1 auxiliary field on $I \setminus b$, holding the reference, basis and pre-step history fixed. This subset fit recomputes its own residual scale, prior-moment precision and raw-basis energy used to normalize the coefficient penalty. Its multiplier remains 0.1. Cap the auxiliary increment at 0.25 times the reference RMS, then sample the fitted field at candidate locations only for sensors in b.

After joining the four held-block predictions, standardize every visible likelihood residual with the shared $\sqrt{v_I}$. Apply the frozen cross-fitted mean and variance for its component mask, multiply by the supplied prior and normalize over sites. The full sensor panel and these weights then determine the final solve. At selection time I is a proposal panel; at final correction it is the full observation panel. Thus subset-fit normalization and likelihood normalization are distinct: the latter is shared across the four folds of I.

A uses one full-panel auxiliary fit with zero standardized mean and unit standardized variance. The calibration-only arm uses that same fit with its separately fitted calibration table. The cross-fit-only arm uses the four held-block predictions with zero mean and unit variance. These operations are the component changes compared in Table 3.

Reference branches and state. Let f_0 be the second complete initial field, n_h the neural prediction from the preceding two reconstructed fields, and P the increment closure. The unshifted reference for branch α is

$$b_h^\alpha = f_0 + \alpha P(n_h - f_0), \qquad r_h^\alpha = b_h^\alpha + Q c_{h-1}^\alpha, \quad \alpha \in \{0, 1/2, 1\}.$$

The four proposal predictions use only their proposal observations in the current neural call. They undergo the same blend and receive the same pre-step shift Qc_{h-1}^α. The orthonormal state basis Q, the branch-specific positive process diagonal d_α and the neural checkpoint are frozen training assets. The state basis is separate from the current correction bank. Initially $c_0^\alpha = 0$ and $V_0^\alpha = 0$; the process covariance is $D_\alpha = \mathrm{diag}(d_\alpha)\max(\mathrm{RMS}(f_0), 10^{-8})^2$.

Prior-moment auxiliary fit. For any fitting subset, recompute v using its reported sites and visible entries. Write M_m as a diagonal visibility matrix, let $\bar r_m = \sum_j \pi_{mj} r(x_{mj})$, and define

$$\Sigma_m^\pi = \sum_j \pi_{mj}(r(x_{mj}) - \bar r_m)(r(x_{mj}) - \bar r_m)^\top,$$

$$W_m = v(M_m \Sigma_m^\pi M_m + vI)^{-1}, \qquad A_m = M_m \sum_j \pi_{mj} B(x_{mj}), \qquad e_m = M_m(y_m - \bar r_m),$$

$$H = \sum_m A_m^\top W_m A_m, \qquad g = \sum_m A_m^\top W_m e_m.$$

Here B contains the closed auxiliary basis columns. "Prior-moment precision" refers to W_m, a variance-scaled working precision. The eight auxiliary coefficients solve $(H_{1:8,1:8} + \lambda I)a = g_{1:8}$, where $\lambda = 0.1 \sum_{k=1}^8 E_k/8$ and E_k is the subset's sampled raw-basis energy defined in Appendix Y. The implemented energy average has floor 10^{-12}. Cap Ba at $0.25\max(\mathrm{RMS}(r), 10^{-8})$ before forming the auxiliary field. This prior-moment fit precedes the location likelihood; the final kernel solve subsequently uses the resulting location weights.

Observation-driven state update. After the current field is returned, every branch updates its own state using the current observations and supplied location prior. Construct H, g, v as above around r_h^α, now using Q as the basis. To penalize accumulated state in weakly observed components, define

$$G = \sum_c (1 - \bar{M}_c) Q_c^\top Q_c, \quad \eta = \frac{\operatorname{tr} H}{\operatorname{tr} G}, \quad H' = H + \eta G, \quad g' = g - \eta G c_{h-1}^\alpha,$$

with $\eta = 0$ when $\operatorname{tr} G \le 10^{-12}$ and $\bar{M}_c$ the observed fraction of component c. Set

$$V^- = V_{h-1}^\alpha + D_\alpha, \quad V^+ = ((V^-)^{-1} + H'/v)^{-1}, \quad \delta = V^+ g'/v.$$

Let d be the number of field scalar entries. Scale δ by $a \le 1$ to keep its norm below $0.25\,\mathrm{RMS}(r_h^\alpha)\sqrt{d}$, then scale $c_{h-1}^\alpha + a\delta$ by $b \le 1$ to keep its norm below $0.5\,\mathrm{RMS}(f_0)\sqrt{d}$. The stored values are $c_h^\alpha = b(c_{h-1}^\alpha + a\delta)$ and $V_h^\alpha = (1-ab)V^- + abV^+$. RMS denominators use floor 10^{-8}. If a component has no visible measurements, the implemented coverage guard instead resets both c_h^α and V_h^α to zero. Each branch is advanced exactly once per observation time, including this bypass.

Calibration histories. For cylinder, the registered kernel-ridge reconstruction generates its own histories on three backbone-training cases. Two cases fit the likelihood parameters; the third checks them without parameter selection. Collection uses H1 and H10 and includes the full fitting panel and four proposal panels with their corresponding references. Calibration for A and cross-fitting is fitted separately. For RBC, trajectories 0 and 1 fit the tables and trajectory 2 supplies the training-side check. These tables are frozen during the reported recurrent experiments; they are not fitted to the evaluated AB trajectories. The true training sites identify residuals for fitting, while reconstruction uses only the supplied supports and priors.

A.4 Fixed-Query Interventions on Calibration

To separate the roles of centering and scale without changing the selector, we replay all 37 uncertain-location queries in the archived configuration-switch panel. Each query retains AA's reference, history and selected configuration. Starting from the frozen mask-conditioned calibration, we either set the fitted mean to zero or set the fitted variance to one; the other parameters are retained. These are parameter-removal interventions, not independently refitted mean-only or scale-only estimators. Each solve recomputes all weight-dependent operators and applies the actual closure and cap. Original A/B outputs are reproduced before interpreting the interventions.

Table 6: Final field NMSE under calibration interventions at fixed queries. Lower is better. No mean retains the fitted scale; no scale retains the fitted mean. Complete retains both. Physical units receive equal weight within each panel/backbone.

Panel	Backbone	Auxiliary fit	No mean	No scale	Complete
Cylinder	FNO	Full	0.001788	0.001752	0.001784
Cylinder	FNO	Cross-fit	0.001804	0.001754	0.001803
Cylinder	LocalNO	Full	0.001136	0.001213	0.001135
Cylinder	LocalNO	Cross-fit	0.001174	0.001191	0.001170
RBC	FNO	Full	0.076599	0.075816	0.076585
RBC	FNO	Cross-fit	0.075400	0.075040	0.075382
RBC	LocalNO	Full	0.052447	0.052239	0.052516
RBC	LocalNO	Cross-fit	0.053254	0.052575	0.053297

Removing scale adjustment changes mean NMSE more than removing centering in all eight panel/backbone/auxiliary-fit groups. Its sign varies: retaining scale helps cylinder/LocalNO, while removing it lowers error in the other three panel/backbone groups. Centering has a smaller effect and can also help or hurt. The panel comprises 14 cylinder/FNO queries, 13 cylinder/LocalNO queries, two RBC/FNO queries and eight RBC/LocalNO queries, selected by the earlier configuration-switch diagnostic. Its means therefore describe that subset. These interventions do not change the complete recurrent component results in Table 3 and do not separate mask conditioning from residual calibration.

A.5 Diagnostics of the Frozen Residual Calibration

We recover the original training-side residual records for all six cylinder case–backbone fits. Recomputing their means and clipped variances reproduces the frozen tables. The saved collector order identifies the physical donor, H1 versus H10 and full versus proposal panels. Of 48 fitted visible-component variance parameters (two auxiliary estimators, six fits and four mask/component entries), 34 reach the lower bound 0.25; none reaches 16. Hidden-component placeholders are excluded. This is parameter-level clipping during fitting, not clipping of a variance re-estimated at every test query.

Table 7 examines cross-fitted residuals at true sites in full-component, full-sensor panels of the training-side held case. After subtracting the frozen mean and dividing by the frozen standard deviation, residual variances still change across the rollout. For example, the calibration associated with case0048/FNO has held variances 0.15/0.13 at H1 and 0.64/0.47 at H10. This residual spread is consistent with the pooled table combining times and fitting-panel sizes. Across the six held panels, standardized component means range from -0.064 to 0.158 at H1 and from -0.159 to 0.006 at H10. The calibrated likelihood is a constrained bias/scale model; the observed field gains do not establish a stationary, independent Gaussian residual law.

Table 7: Training-side held-case diagnostics. Each row identifies the evaluated model's outer fold; residuals come from its held training donor. Variances are for standardized u/v residuals; correlation uses jointly visible residuals.

Outer case	Backbone	Variance H1	Variance H10	Corr. H1	Corr. H10
0019	FNO	0.30/0.34	0.49/0.36	-0.04	-0.01
0019	LocalNO	0.21/0.27	1.85/0.90	0.04	0.20
0043	FNO	0.33/0.46	1.32/1.80	0.05	-0.22
0043	LocalNO	0.58/0.58	2.11/1.65	0.07	0.07
0048	FNO	0.15/0.13	0.64/0.47	-0.09	0.14
0048	LocalNO	0.11/0.14	0.37/0.51	-0.04	0.13

RBC. No visible-component variance parameter reaches a clipping bound. On the already evaluated reserved trajectory 30, full-panel cross-fitted standardized variances are FNO: 0.28/0.90 at H1 and 0.92/1.57 at H10; LocalNO: 0.79/1.16 at H1 and 1.16/0.90 at H10. The absence of clipping separates this panel from the cylinder lower-bound effect. Temporal variation remains, but these residual summaries alone do not identify the cause of negative field gain. Fitting still uses training trajectories 0/1 and checking uses trajectory 2.

A.6 A Fixed-Scale Alternative to Learned Calibration

The variance-floor diagnostic motivates a simple alternative: set every visible normalized residual mean to zero and every variance to 0.25, without fitting separate statistics by mask. We apply this rule to both the full auxiliary fit and the cross-fitted auxiliary fit. The weights become $q_{mj} \propto \pi_{mj} \exp(-2 \sum_{c \in M_m} e_{mjc}^2)$, where e retains the reference-based residual normalization. Thus the rule sharpens the likelihood while leaving the location prior unchanged.

Both alternatives use A for selection on their own histories and change only the final location weights. The subsequent observation operator, kernel solve, closure and cap are recomputed as usual. We freeze three cylinder cases, two backbones, observation batch 17100, 64 sensors and all eight conditions over H1–20. The original AA, cross-fitting-only, calibration-only and AB results use the same observations. Each of the twelve new tasks reproduces its corresponding learned estimator at H1 before the intervention; all 1,920 new rows pass the observation and completion checks.

Table 8: Full independent recurrence with learned or fixed likelihood statistics. CF denotes cross-fitting without calibration; Cal uses learned calibration with the full auxiliary fit. Fixed/full and Fixed/CF use zero means and variance 0.25 with their respective auxiliary fits. Lower NRMSE is better.

Case	Backbone	AA	CF	Cal	AB	Fixed/full	Fixed/CF
0019	FNO	0.011336	0.011256	0.010981	0.010917	0.010974	0.010899
0043	FNO	0.018192	0.018146	0.017679	0.017136	0.017674	0.017487
0048	FNO	0.034083	0.034799	0.033605	0.034105	0.033603	0.035586
Mean	FNO	0.021204	0.021400	0.020755	0.020719	0.020750	0.021324
0019	LocalNO	0.011193	0.011113	0.010974	0.010914	0.010974	0.010899
0043	LocalNO	0.017483	0.015942	0.015937	0.016677	0.015944	0.016609
0048	LocalNO	0.034724	0.034880	0.034352	0.034773	0.034056	0.035388
Mean	LocalNO	0.021133	0.020645	0.020421	0.020788	0.020325	0.020965

With the full auxiliary fit, fixed sharpening improves over AA by 2.14%/3.83% for FNO/LocalNO, slightly exceeding learned calibration by 0.02%/0.47%. With cross-fitting, learned statistics reduce error relative to the fixed scale by 2.83%/0.85%, but win only two FNO cases and one LocalNO case. Compared directly with the simpler full-fit fixed-scale alternative, AB improves FNO by 0.15% and worsens LocalNO by 2.28%. The benefit of learned statistics within the cross-fitted estimator therefore does not establish that the complete AB design is necessary.

This is a post-hoc diagnostic on previously evaluated cases and observations. The fixed variance was motivated by the fitted variance floor, not selected through a new temperature search. The comparison tests this concrete simple substitute; it does not isolate mask conditioning from centering and component-specific scaling, or establish performance against an optimally tuned global temperature.

B Public-Architecture Comparison and Provenance

B.1 Comparator identities and information

The RecFNO core is imported from the authors' repository at commit `93c71db4826314c752ac9a63ad00ec541e8f9b5c`. Its source matches that commit after normalizing Windows line endings. The pre-existing adapter concatenates three eight-channel inputs (two history fields and a nearest-visible current field), uses two output components, and initializes a residual output head at zero during training. These are changes to the original task interface and output parameterization; they are disclosed rather than described as an unchanged end-to-end implementation.

Training uses three cases per fold, two separate calibration cases and one excluded outer case. The original 11 checkpoint looks and calibration selections remain fixed. No neural reference, auxiliary estimates or dense current target are supplied to the comparator. Both programs start with two complete initial fields and thereafter use their own outputs; RecFNO receives its own previous/current fields. Native outputs are the primary comparison. Postclosed outputs apply the common closure to those same native trajectories without feeding the closed field back.

There are 480 unique case–condition–time queries and 960 pairings with the two AB backbones, not 960 independent physical samples. Observation hashes match both AB archives. The evaluation checks checkpoint and training-data provenance, recomputes CAL choices, tests invariance to hidden-component NaNs, and reproduces AB at the first step. One bookkeeping repair canonicalized integer checkpoint keys after JSON serialization; no numerical method or selection was changed.

FLRONet source and checks. The second comparator uses the authors' `FLRONetFNO` core at commit `f9026bf4405ba4b7af902be9f5851973ef5983ae`. Its historical adapter provides three eight-channel time slots with the same causal history and visible-value allowance, and a residual output initialized at zero. The core is unchanged; the causal interface and output parameterization are task adaptations. Checkpoint selection is recomputed from all 11 original CAL looks. The paired evaluation independently passes all 480 query fingerprints, checkpoint/source checks, hidden-value

Table 9: Source, adaptation, fitting information and scope of the comparisons. No row is labeled an original-benchmark official reproduction. Public cores and project combinations answer different questions.

Identity/source	Adaptation or shared machinery	Training and calibration	Evaluation and interpretation
RecFNO / FLRONet: pinned author cores	Causal vector/history, masks, geometry and horizon inputs; residual output. Native feedback, no AB kernel or selector.	Per fold: three train, two CAL, one outer case; two RecFNO seeds, one FLRONet seed; 4,800 updates; 11 CAL checkpoint looks.	Three cylinder cases, eight conditions, H1–20. Clean is the common-information anchor; corruption extensions differ in explicit location-law use.
Senseiver / PalSB: pinned author cores	Sparse-vector history interfaces; own recurrent outputs. PalSB retains the bridge and sampled-path physical finetuning with task geometry.	Same cylinder train/CAL splits. Senseiver: 4,800 updates/11 looks. PalSB: two pretraining starts, 1,000 fine-tuning updates each, six CAL candidates.	Three cylinder cases, eight conditions, H1–20. Selected PalSB checkpoints precede finetuning. Details in Appendix B.4.
AB: current project system	A selects reference/configuration; cross-fitted, calibrated B supplies final kernel weights; own recurrence.	Frozen pretrained references and training-fitted auxiliary calibration. Same common-program candidate opportunities as below.	All current panels. External-system gains do not isolate one component.
A–R / R–R: project reference alternatives	Same kernel, references, postprocessing and 109 settings. Reference weights replace final handling only / both stages.	Reference residuals independently fitted on the same training donors with the auxiliary calibration family; parameters frozen across windows.	Three cylinder cases, two backbones and windows, eight conditions, H1–20; incremental auxiliary attribution.
GP moments: project common-program comparator	Prescribed location moments, including shared sensor location; common solver, scoring and candidate grid.	Shares frozen references/backgrounds; no auxiliary residual calibrator. Hyperparameters selected by the common sensor-held-out rule.	Current recurrent panels; a matched inference alternative, not official NIGP reproduction.
AA / BB and single-component variants: internal	Original / modified weights in both stages; cross-fit-only and calibration-only retain their registered assignments.	Same references and candidate opportunities; required training calibration frozen before evaluation.	Stage and component ablations. BB wins and negative conditions remain reported.
NIGP-inspired: project adaptation	Shared background with input-gradient covariance; its own optimization/selection hierarchy.	Frozen reference and adaptation-specific fitting, detailed in Appendix M.	Eight 3D spatial objects, first windows; supporting sensitivity, not pooled with the common-GP column.

invariance and the AB first-step replay. Restoring missing remote assets and registered line endings changed no numerical code. These are the same three physical cases as the RecFNO comparison, not additional physical samples.

Table 10: Complete recurrent NRMSE on three cylinder cases, averaged equally over cases. Both public architectures use frozen task adapters and native feedback. The postclosure sensitivity is reported in Appendix B. The primary anchor is accurate locations with all components (clean). The primary table uses one seed per fold, 4,800 updates and 11 CAL looks; training/CAL and interface details are in Table 9.

Condition	RecFNO author-core adapter	FLRONet author-core adapter	AB/FNO current system	AB/LocalNO current system
Clean	.10657	.13959	.01316	.01234
Location jitter	.10833	.14039	.02533	.02506
Missing u	.11534	.14071	.01580	.01403
Jitter, missing v	.18455	.14706	.02756	.02999
All eight conditions	.12913	.14233	.02112	.02051

Table 11: Clean-condition current-system comparison, H1–20 NRMSE.

Case	RecFNO native	Postclosed	AB/FNO	AB/LocalNO
0019	.167916	.163312	.007733	.007733
0043	.045023	.043698	.011747	.008649
0048	.106767	.093905	.019995	.020641

Across all eight conditions, postclosed mean NRMSE is 0.123578 for RecFNO and 0.133752 for FLRONet, compared with native means 0.129133 and 0.142327. No output variant is selected using these test errors. Finite training budgets and early/boundary CAL optima leave convergence sensitivity unresolved for both architectures.

Bounded RecFNO optimization sensitivity. For case 0019, whose original CAL choice reached the training budget, we resumed the saved step-4,800 model, optimizer and random states for 2,400 additional updates. The continuation retains the original final learning rate 10^{-4} and extends the optimization budget. Two additional CAL looks at 6,000 and 7,200 compete with the original eleven looks. CAL selects step 7,200, improving its score by 5.24%. This choice is sealed before outer evaluation. All 160 outer query fingerprints and 320 AB pairings pass the evaluation audit.

The three-case all-condition error falls by 1.31%; AB reductions remain 83.43%/83.90% for FNO/LocalNO. The original-budget main table and this sensitivity are reported separately. This first continuation selected its endpoint. A further bounded check, below, tests whether continued fitting materially changes the comparison; neither check adds a physical case.

B.2 RecFNO training-seed repeat

An additional training seed (832 versus 831) retains the original architecture, training observations, 4,800-update budget and 11 CAL checkpoint looks. All three checkpoint choices were sealed before outer evaluation; the selected steps are 1,600/1,600/50 for cases 0019/0043/0048. The outer observation bank remains 5100, separate from the five-arm seed-17100 experiment. All 480 queries and 960 backbone pairs pass the artifact audit. The native three-case NRMSE is 0.122764 over all conditions and 0.108156 on clean observations, compared with 0.129133 and 0.106569 for the original seed. AB reductions against the new seed are 82.80%/83.29% over all conditions and 87.83%/88.59% on clean observations for FNO/LocalNO. This repeat supports robustness to this training-seed change within the tested adapter and budget; it adds no physical cases and does not establish optimization convergence.

Historical NODA results use a paper-description sparse-vector reimplementation and a different observation batch. They are not current AB results or official-code reproduction. The common-base GP moments and the NIGP-inspired adaptation are separately identified in the main text and Appendix M. No result from the unfinished joint-location GP sampling comparison is substituted for these methods.

Table 12: FLRONet task adaptation on clean conditions, H1–20 NRMSE. Postclosed outputs retain native feedback.

Case	Native	Postclosed
0019	0.162930	0.137220
0043	0.060926	0.060537
0048	0.194927	0.194914

Table 13: RecFNO native-output sensitivity. Only the 0019 fold is continued; the other two folds and all AB results are unchanged. Lower NRMSE is better.

Scope	Conditions	Original budget	With continuation
0019	Clean	0.167916	0.165126
0019	All eight	0.177734	0.172663
Three cases	Clean	0.106569	0.105638
Three cases	All eight	0.129133	0.127443

B.3 Training Exposure and Early Calibration Optima

The RecFNO and FLRONet comparators and the FNO/LocalNO references are trained from fresh weights on the same three training cases per fold, excluding two calibration cases and the outer case. The recurrent neural objectives use H1–20 with detached feedback, four corruption conditions and alternating 64/256 sensors. Equal update counts align exposure, not FLOPs or parameter counts.

Table 14: Original neural fitting budgets and subsequent frozen-system fitting. New continuation checks retain the original comparisons as their reference.

Program	Neural updates / CAL looks	Additional training-side information
AB: FNO or LocalNO	4800 / 11	Training-derived background/state banks and process diagonals; two training cases fit auxiliary likelihoods, the third checks them. At inference, supplied location priors enter explicitly.
RecFNO adaptation	4800 / 11	Same three neural training cases; second training seed evaluated. Fold0 continuation starts from saved model, optimizer and RNG states.
FLRONet adaptation	4800 / 11	Same three neural training cases; one training seed. Fold4 receives the bounded continuation check.

The kernel and auxiliary stages give AB additional structure and training-derived assets; the public-architecture comparison measures this complete program. The independently calibrated reference and component comparisons control these assets when attributing the location estimator.

For the fold whose outer case is 0048, both public adaptations select an early CAL checkpoint. Four fixed training-only recurrent minibatches clarify this behavior: RecFNO mean training NMSE falls from 0.01403 at step50 to 0.00005145 at step4800, while its CAL score rises from 0.5212 to 0.9210. FLRONet training NMSE falls from 0.06144 at step100 to 0.0005388 at step4800, while CAL rises from 0.3926 to 1.1460. This supports continued training fit with deteriorating calibration performance. The probe does not establish optimization convergence or exclude gains from a different training procedure.

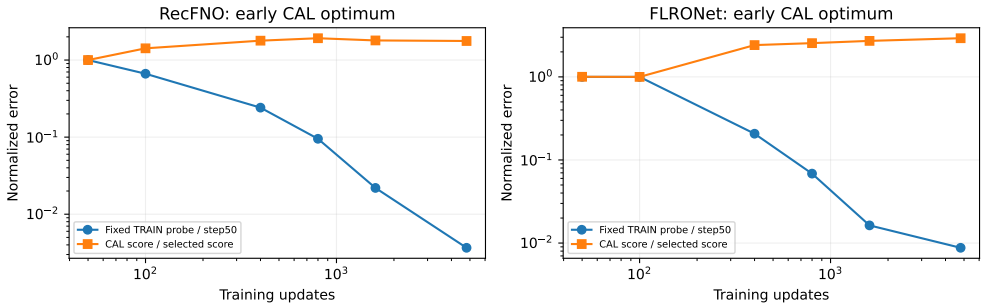

Figure 3: Fixed training probes and CAL errors for the early-optimum fold. Training curves are normalized to step50; CAL curves to their selected checkpoint. The separate normalizations compare trends, not absolute train–CAL error gaps.

A second bounded RecFNO continuation. The case0019 model resumes its saved step7200 state to step9600 at the same 10^{-4} learning rate. Two further CAL looks select step8400 (0.881745), while step9600 scores 0.905300; the previous step7200 score was 0.903475. The selected checkpoint's clean NRMSE is 0.165481 native and 0.161057 postclosed, and its all-condition NRMSE is 0.172091 native and 0.167762 postclosed. Thus the latest selected model leaves the large gap to the complete AB system essentially unchanged. The optimum is no longer the final inspected checkpoint, although this bounded check does not establish convergence.

B.4 Senseiver and PalSB task adaptations

We also evaluate author-core adaptations of Senseiver (Santos et al., 2023) and PalSB (Li et al., 2025) on the same three cylinder cases. Each uses its own reconstructed fields as recurrent history. Checkpoints are selected on the two calibration cases excluded from training, and all 480 outer queries are paired with both AB backbones. Native outputs are reported in Table 15; these results add comparators, not physical cases.

Table 15: Additional public-architecture adaptations: native H1–20 NRMSE. Means weight the three cylinder cases equally. Senseiver uses its original 4,800-update budget. PalSB selects from the registered pretraining and physical-finetuning candidates described below.

	Senseiver		PalSB	
Case	Clean	All eight	Clean	All eight
0019	0.269160	0.275902	0.750784	0.777961
0043	0.071181	0.126086	0.074669	0.137691
0048	0.177900	0.199596	0.077858	0.135092
Mean	0.172747	0.200528	0.301104	0.350248

Senseiver fitting. The adapter reconstructs signed vector fields from the two initial/history fields, visible sensor values, component masks and reported sites. It receives no AB auxiliary field, kernel solver or explicit position prior. Training uses the same three cases, 4,800 recurrent updates and eleven CAL checkpoint opportunities as the original external comparison. CAL selects steps 4,800, 4,800 and 800 for cases 0019, 0043 and 0048. The first two curves improve at the budget endpoint; Table 15 therefore identifies a fixed-budget comparison. For 0048, the selected earlier checkpoint precedes deterioration in the aggregate CAL score.

Senseiver bounded continuation. Cases 0019 and 0043 resume the saved model, optimizer and random states from step 4,800 for 2,400 additional updates, retaining the final learning rate 10^{-4}. Two new CAL looks at steps 6,000 and 7,200 compete with the original eleven. Both folds select step 7,200; case 0048 retains its original step-800 model. All 320 new outer queries pass the observation and AB-pairing checks. Table 16 separates the original budget from this extension.

Table 16: Senseiver budget sensitivity: native H1–20 NRMSE. The extension changes two folds; case 0048 is unchanged. Means weight the three cases equally.

	Original budget		With continuation	
Case	Clean	All eight	Clean	All eight
0019	0.269160	0.275902	0.266391	0.278174
0043	0.071181	0.126086	0.064468	0.135329
0048	0.177900	0.199596	0.177900	0.199596
Mean	0.172747	0.200528	0.169587	0.204366

Continuation lowers the three-case clean mean from 0.172747 to 0.169587, but raises the all-condition mean from 0.200528 to 0.204366. Better CAL scores therefore do not translate into improvement across all outer observation conditions. The continued models are still selected at their budget endpoints, so this check does not establish convergence. The complete AB system retains lower clean error in each case.

PalSB physical finetuning. The adapter retains the author's conditional UNet and I2SB bridge, with two-component inputs, training-only channel normalization and 100-step inference. Its task pretraining uses 24,000 bridge-matching updates. We supplement the original single-step penalty adaptation with the author's ten-step sampled trajectory and truncated backpropagation: the implemented $K_{\mathrm{sg}} = 1$ retains gradients through the final two network calls. The loss weights are 10 for normalized-field L1 matching and 0.5 for physical-field divergence energy normalized by field energy. AdamW uses learning rate 10^{-5} and no weight decay; StepLR multiplies the rate by 0.99 every ten updates. Each branch takes 1,000 updates with four-sample subsets of 32-sample training batches. Geometry-specific spacings and valid fluid stencils adapt the author's cylinder divergence loss to our grids; the old extra squared-grid-spacing attenuation is removed.

Two predeclared starting points are tested: the 24,000-step pretraining endpoint and the previously CAL-selected 8,000-step checkpoint. CAL compares 0, 500 and 1,000 finetuning updates from each, giving six candidate looks. All three folds retain the unfinetuned 8,000-step checkpoint. Finetuning from 24,000 steps improves two folds relative to that starting point, but does not beat the earlier selected model. Thus the final errors remain unchanged after giving the rollout-based physical phase an opportunity to improve the selection. The selected models in Table 15 are pre-finetuning checkpoints; the table is not evidence of a finetuning gain.

The implementation check reproduces the author's sampled-path predictions and parameter gradients exactly on a training minibatch. The cylinder divergence stencil also agrees on an equal-spacing full-fluid grid. The task retains its own observation interface, geometry and pretraining budget; unused outer reward monitoring is omitted. These are author-core task adaptations rather than original-benchmark reproductions.

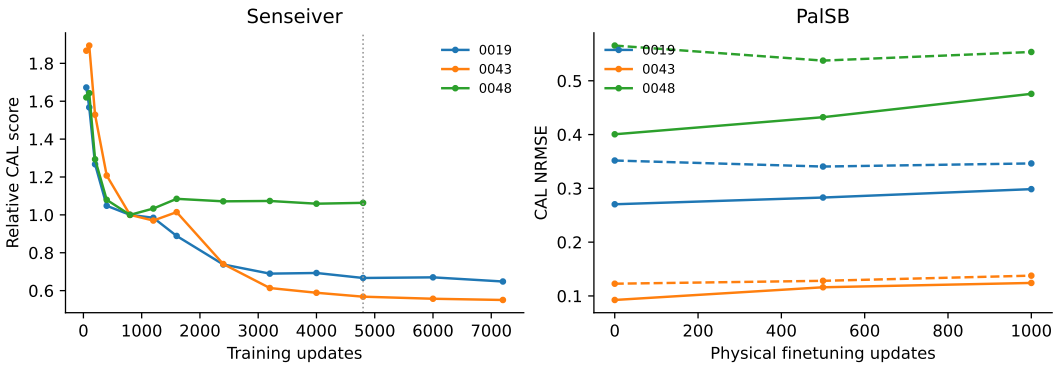

Figure 4: Training-side selection evidence. Left: Senseiver CAL scores normalized by the original step800 checkpoint, including the two continued folds; the vertical line marks the original budget. Right: PalSB CAL NRMSE under the author-rollout finetuning from two predetermined starts; solid lines start at 8,000 pretraining updates and dashed lines at 24,000. Colors identify the held-out case, not the calibration data.

FLRONet bounded continuation. For case 0043, 2,400 further updates retain the original final learning rate 10^{-5} and the saved optimizer and random states. The original eleven CAL looks compete with steps 6,000 and 7,200. CAL selects step 6,000: its score is 0.698571, versus 0.710752 at 4,800 and 0.703948 at 7,200. The selected clean NRMSE is 0.059469 native and 0.059147 postclosed; all-condition NRMSE is 0.066254 native and 0.065958 postclosed. The native all-condition error falls by 3.35% from 0.068551. The original-budget table remains separate from this continuation; the selected model improves without changing the complete-system ranking.

C Independently Calibrated Reference Alternatives

Six reference-residual calibration models use the same training trajectories, donor subsets, mask classes and mean/variance estimation family as auxiliary calibration. They are fitted separately: auxiliary parameters are not reused for reference residuals. Outer cases are excluded from the associated training and calibration. Parameters are unchanged in the later window. Twenty-four complete tasks cover two reference alternatives, two windows, three cases and two backbones. Each task has eight conditions and twenty recurrent steps. The resulting 3,840 paired method rows are repeated evaluations of three physical cases.

Table 17: AB relative NRMSE reduction over R–R, with the reference alternative allowed to select its own configuration. Positive favors AB.

	Original window		Later window	
Case	FNO	LocalNO	FNO	LocalNO
0019	−0.27%	1.06%	−2.16%	0.82%
0043	1.28%	−1.22%	−4.06%	−6.84%
0048	13.73%	14.88%	13.45%	12.32%

R–R was introduced after inspecting the original A–R comparison; this chronology is part of the protocol. It was evaluated on all registered case–backbone groups in both windows, without retuning AB or reference calibration. Both comparison arms share the predictive-reference machinery and background basis. R–R removes the auxiliary field from location-weight generation, not every inherited state-estimation component. Gains against R–R cannot be presented as gains against the stronger common-base GP.

D Directional Precision Control

D.1 Directional error and precision control

For the fixed production systems in Eq. (6), write $S_A = LL^\top$, $w_A = S_A^{-1} e_A$, and

$$z = e_B - e_A - (S_B - S_A)w_A, \quad g = L^{-1}z, \quad M = L^{-1}S_B L^{-\top}. \tag{12}$$

Then $w_B = w_A + L^{-\top}M^{-1}g$. With $E = M - I$, the second-order inverse approximation uses $x_2 = (I - E)g$ and has exact remainder

$$M^{-1}g - x_2 = (I + E)^{-1}E^2 g. \tag{13}$$

In an eigenvector of E, its coefficient is multiplied by $\lambda_i^2/(1 + \lambda_i)$ before propagation through $D_B L^{-\top}$ and closure. Thus a scalar perturbation norm alone omits both residual excitation and field propagation. The equality follows by multiplying both sides by $I + E$; it does not require a convergent Neumann series.

Postprocessed response bound. For any approximation x to $M^{-1}g$, let $r_x = Mx - g$. With the reference and its cap radius fixed, the cap is a Euclidean ball projection and is nonexpansive in the full-grid norm. Closure need not be orthogonal. Applying its actual linear operator P gives

$$\|F(x) - F(x_*)\|_{\text{fluid}} \le \frac{\|PD_B L^{-\top}\|_F}{\lambda_{\min}(M)} \|r_x\|_2 = \varepsilon_x. \tag{14}$$

Indeed, $x - x_* = M^{-1}r_x$, the field error before capping is $PD_B L^{-\top}M^{-1}r_x$, and restricting to fluid points cannot increase the full-grid Euclidean norm. The Frobenius norm upper-bounds the operator norm. These statements assume the positive-definite working systems and fixed reference used in this experiment; they do not cover reference reselection.

If $a = \|F(x) - F_A\|_{\text{fluid}} > \varepsilon_x$, the response relative error is at most $\varepsilon_x/(a - \varepsilon_x)$ and its cosine is at least $\sqrt{1 - (\varepsilon_x/a)^2}$. The first follows from the reverse triangle inequality; the second from the tangent to the error ball centered at the predicted response. These bounds use observations and working operators, without the target field or the exact B solution.

We evaluate conjugate gradients at registered checkpoints up to 32 iterations and stop when these sufficient conditions meet the fixed precision criteria. A scale guard additionally requires $a - \varepsilon_x > 10^{-8}\max\{1, \|F_A\|_{\text{fluid}}\}$ and otherwise withholds relative certification. This is a numerical use of standard inverse and residual bounds; the experiment tests their usefulness for the particular production response, rather than proposing a new Krylov solver. Floating-point evaluation is not an interval-arithmetic proof. Equation (10) remains the separate test of reconstruction benefit.

E Three-Dimensional Current-Method Evaluation

E.1 Current-method three-dimensional evaluation

The frozen correction methods were evaluated in 108 tasks on JHTDB: six MHD spatial objects with two starting windows each (H1–6), and two isotropic-turbulence objects with three windows each (H1–20), using two backbone families and three reference-model seeds. All methods share observations, geometry and candidate opportunities and retain independent histories. The reference predictors include a historical frozen sparse posterior; they are not bare FNO or LocalNO. These data sources have prior project exposure. MHD favors AB over all three correction comparators in the three uncertain-location condition means. Isotropic turbulence favors AA and GP moments over AB; AB is only slightly better than BB. Accurate-location correction controls coincide. Extending the same design to 3D therefore preserves conditional gains and failures rather than establishing universal stage-separation benefits.

The 18 windows are not 18 independent physical objects. In particular, 44 groups have matching metric sequences across repeated model tasks, concentrated in MHD, where almost all correction outputs select $\alpha = 0$. Such repetitions do not supply independent evidence of seed robustness. Neither the dataset labels nor these aggregate rankings identify the physical cause of the difference.

Table 18: Current 3D NRMSE, averaged within each window over horizons, backbones and seeds, then equally over windows. Lower is better. GP denotes the common-base moment adaptation.

Panel	Condition	AA	AB	BB	GP
MHD	Jitter	.24072	.23694	.23932	.24196
MHD	Jitter/missing z	.24428	.23750	.24309	.24693
MHD	Jitter/noise	.24020	.23672	.23832	.24082
Isotropic	Jitter	.19372	.19507	.19524	.19146
Isotropic	Jitter/missing z	.19426	.19543	.19585	.19475
Isotropic	Jitter/noise	.19390	.19503	.19517	.19163

Fixed-history diagnostic. A post-hoc replay uses the first registered window and smallest seed of each spatial object and backbone, retaining the logged AA reference, configuration and history. All 16 tasks reproduce AA metrics within the fixed tolerance; isotropic neural predictions required the original GPU path after CPU replay exceeded that tolerance. Changing only final A/B weights already harms the isotropic means in all three uncertain-location conditions: mean NMSE changes are $+0.00057$, $+0.00050$ and $+0.00064$ for jitter, jitter/missing and jitter/noise. Mean location NLL also worsens. The corresponding MHD changes are -0.00214, -0.00341 and -0.00147, with improved mean location NLL. Isotropic failure is therefore present before changing selection or recurrent history.

The isotropic risk cross-term is positive in each condition mean. In a shared reference-NRMSE stratum $[0.2, 0.3)$, mean reference errors are 0.2444/0.2434 for MHD/isotropic, yet local loss changes retain opposite signs. Excessive magnitude or reference error size alone therefore does not explain the difference. This target-dependent diagnostic does not yield an observation-only deployment rule.

Additional GP route. A shared-background, vector NIGP-inspired adaptation was evaluated on the first window of all eight objects, with both backbones and one seed (Appendix M). AB has 0.78–1.62% higher uncertain-location mean NRMSE on MHD and 0.40–1.66% lower error on isotropic. The adaptation retains the uncorrected reference on 346/360 MHD steps, so its advantage does not establish better GP corrections. It uses continuous likelihood fitting and a different selection hierarchy; this complete-procedure comparison is distinct from the common-grid moment comparison.

F Additional Component Interpretation

F.1 Component effects at fixed references

The six-rule controlled comparison holds the reference, history and full kernel configuration fixed across 96 cylinder queries. Cross-fitting improves all six case–backbone means, while calibration of the original estimator improves three; their combination improves all six means over either single change. Reselection changes these rankings in three groups. Appendix O gives the component table and intervention protocol. These development interventions motivate the independent recurrent ablation below rather than replace it.

F.2 Component effects in independent recurrence

Table 19 gives the original observation-batch case-level ablation. The completed additional seed is reported below. We complete the component comparison by adding two recurrent arms to the cylinder confirmation panel: A selects the reference and configuration, while either cross-fitting alone or calibration alone generates the final weights. Each arm updates its own field history and coefficient states. The twelve additional tasks contribute 1,920 rows paired with the archived AA and AB runs at 960 common queries.

Both single changes improve the mean over AA in all six case–backbone groups. Their combination gives the lowest aggregate mean for each backbone: the additional reduction over cross-fitting alone and calibration alone is 4.07% and 1.23% with FNO, and 0.22% and 0.42% with LocalNO. The combination is case dependent. Calibration alone is better than AB in case 0019 with either

Table 19: Complete H1–20 component ablation on the original cylinder observation batch (seed offset 5100). All arms use A for selection on their own histories. Values are relative NRMSE reductions over AA; positive favors the modification.

Case	Backbone	Cross-fit only	Calibration only	Both (AB)
0019	FNO	0.90%	2.71%	1.96%
0019	LocalNO	0.89%	2.80%	2.30%
0043	FNO	2.20%	4.84%	7.62%
0043	LocalNO	1.03%	4.15%	2.68%
0048	FNO	0.84%	4.12%	5.10%
0048	LocalNO	4.82%	2.57%	4.05%
All cases	FNO	1.23%	4.07%	5.25%
All cases	LocalNO	3.19%	2.99%	3.40%

backbone and in case 0043 with LocalNO; cross-fitting alone is better in case 0048 with LocalNO. AB outperforms both single changes in the remaining two groups.

The differences also vary over the rollout. With LocalNO, AB is better than either single component over H1–5, but has 1.66% higher mean error than either over H11–20. Condition-level harm remains: cross-fitting alone increases error by 11.24% in case 0048/FNO under `combined-u25`. This original-batch experiment supports complementary aggregate contributions, with case- and horizon-dependent exceptions. The AB definition remains fixed across cases.

F.3 Completed bounded repeats

The five-arm repeat changes the observation-seed offset from 5100 to 17100, retaining the same three physical cases, window, frozen predictors, calibration and candidate opportunities. All six tasks complete H1–20 over eight conditions with independent histories (960 case–backbone queries). This tests observation randomness, not additional physical generalization or backbone-training seeds. The component extension adds twelve tasks and 1,920 rows on the same bank. Its scope was frozen before computing the new component arms, but after inspecting early five-arm results; it is not an untouched pre-access confirmation. All twelve component implementation replays and six five-arm replays passed before evaluation.

Table 20: Completed additional observation seed: component reductions over AA (percent; positive is better). Original-batch values remain in Table 19.

Case	Backbone	Cross-fit only	Calibration only	AB
0019	FNO	0.70	3.13	3.70
0019	LocalNO	0.71	1.96	2.49
0043	FNO	0.26	2.82	5.81
0043	LocalNO	8.82	8.84	4.61
0048	FNO	-2.10	1.40	-0.06
0048	LocalNO	-0.45	1.07	-0.14
All	FNO	-0.93	2.12	2.29
All	LocalNO	2.31	3.37	1.63

Calibration improves all six repeated case–backbone means; cross-fitting improves four. AB outperforms both single changes in three groups, versus two in the original batch. At the aggregate level, AB improves over calibration alone by 0.17% with FNO, but has 1.80% higher error with LocalNO. Thus repeatable calibration benefits do not establish that both components are necessary in every setting.

Table 21: Five-arm repeat: aggregate reductions of frozen AB (percent). R–R has independently trained reference-residual calibration.

Cases	Backbone	vs. AA	vs. BB	vs. GP	vs. R–R
All three	FNO	2.29	0.60	0.94	8.30
All three	LocalNO	1.63	-2.85	-1.88	3.34
Without 0048	FNO	5.00	0.84	5.38	-0.41
Without 0048	LocalNO	3.78	-5.57	-1.06	-7.54

Average gains over R–R repeat, but excluding 0048 reverses this comparison. In contrast, gains over AA occur mainly in 0019 and 0043. These are different attribution questions: the former tests auxiliary inference against reference calibration, and the latter tests the modified estimator against the original auxiliary procedure. GP and BB comparisons remain mixed. The repeat neither changes the definition of Ours nor introduces a per-case best-arm selector.

F.4 Physical-Unit Robustness and the 0048 Effect

We reaggregate the completed archives by physical case before comparing methods. Each case has equal weight; backbones and observation batches remain separate. Table 22 reports the ratio of case-mean NRMSEs and its range after removing each of the three cases. These ranges are sensitivity summaries, not confidence intervals. The two observation batches reuse the same physical objects.

Table 22: AB reduction in NRMSE (percent); brackets give the leave-one-case range. Positive favors AB.

Backbone	Comparator	Original observations	Repeated observations
FNO	AA	5.25 [4.35, 5.94]	2.29 [0.87, 5.00]
FNO	BB	1.51 [-0.20, 4.68]	0.60 [0.04, 1.09]
FNO	Calibration only	1.23 [0.59, 1.65]	0.17 [-0.98, 2.12]
FNO	GP moments	2.28 [-0.08, 5.65]	0.94 [-1.75, 5.38]
FNO	R–R	8.34 [0.67, 10.68]	8.30 [-0.41, 11.43]
LocalNO	AA	3.40 [2.52, 3.65]	1.63 [0.50, 3.78]
LocalNO	BB	0.71 [0.17, 2.42]	-2.85 [-5.57, 0.13]
LocalNO	Calibration only	0.42 [-1.09, 1.04]	-1.80 [-2.53, -0.80]
LocalNO	GP moments	1.22 [-1.37, 6.06]	-1.88 [-2.76, -1.06]
LocalNO	R–R	9.03 [-0.23, 11.89]	3.34 [-7.54, 8.43]

Removing any single case preserves a positive aggregate AB–AA effect in both backbones and observation batches. Comparisons against calibrated reference inference, GP moments and the single-component variants are more sensitive. In the later time window, AB's small aggregate advantage over GP moments has leave-one-case ranges of $[-0.95, 1.98]\%$ with FNO and $[-0.55, 1.77]\%$ with LocalNO. That window changes both time and observations and is not pooled with the original-window seed repeat.

Where the 0048 benefit occurs. Against R–R, AB improves the mean in all 18 combinations of backbone, time bin (H1–5, H6–10, H11–20) and evaluation batch (original, repeated observations, later window). Against GP moments, it improves only five of these combinations. Thus its advantage over calibrated reference weights persists across time bins but does not establish superiority to the moment alternative. Table 23 shows the repeated observation batch. Accurate-location conditions yield identical fields for these weight alternatives. All four uncertain-location conditions favor AB over R–R in this batch; reductions range from 14.73–28.91% with FNO and 7.57–23.36% with LocalNO. These independent-rollout error differences alone do not establish a sustained improvement in location probability accuracy.

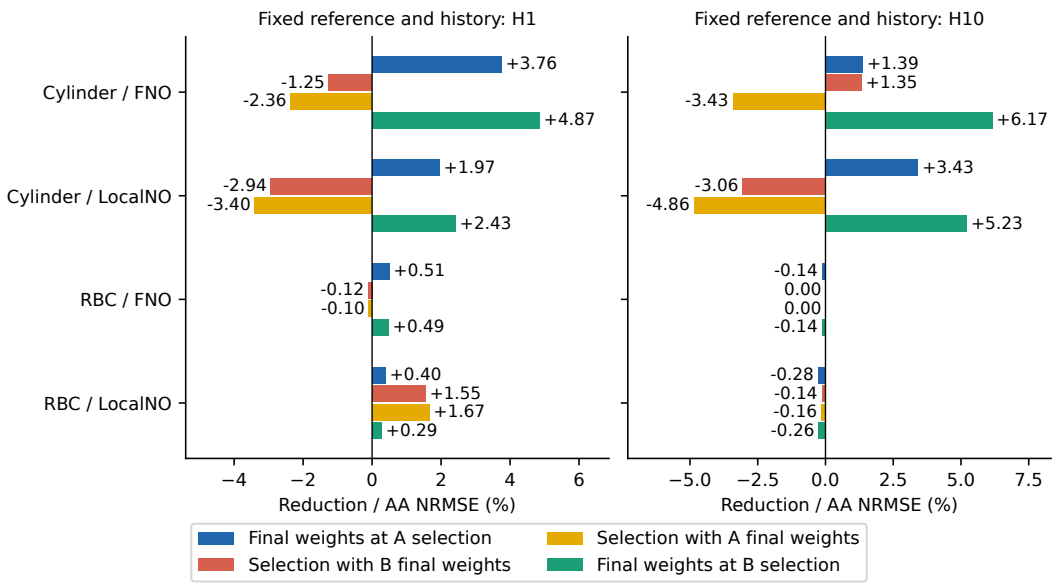

Figure 5: Two paths through the conditional four-way comparison with the AA reference and history fixed. Blue then red follows AA–AB–BB; yellow then green follows AA–BA–BB. All reductions use the same AA NRMSE denominator, so both paths have the same total. The individual effects differ across paths, backbones and horizons.

Table 23: Case0048, repeated observations: AB reduction (percent) over two alternatives, within each time bin.

Backbone	Time	vs. R–R	vs. GP moments
FNO	H1–5	15.44	1.12
FNO	H6–10	20.19	-4.25
FNO	H11–20	9.72	-4.00
LocalNO	H1–5	13.40	-1.13
LocalNO	H6–10	13.82	-4.06
LocalNO	H11–20	7.09	-2.15

G Detailed Stage Interventions

G.1 What changes when location weights enter selection?

At H1 and H10 of the AA trajectory, we retain the reference, history and observations and recover the conditional AA/AB/BA/BB table from saved candidate risks and four-panel scores. All 176 queries pass 792 reconstruction, selection and identity checks. This intervention changes the configuration within one reference. Independent recurrence comparisons additionally include reference choices and accumulated history. Appendix T reports the earlier complete four-arm development evaluation, kept separate from this new-observation panel.

At H1, changing final weights improves mean cylinder NRMSE for both backbones, while the subsequent configuration change reduces that benefit. The direction is not universal: at H10 the FNO mean benefits from reselection, whereas LocalNO loses much of its fixed-configuration improvement. For RBC, the fixed-reference benefit of changing weights at H1 becomes a small average loss at H10 in both backbones. The preferred stage assignment changes across backbones and horizons.

The reverse path separates stage effects at the panel and query levels. On cylinder/FNO at H10, replacing A selection by B selection increases mean error when the final estimator is A (a -3.43% reduction), but reduces it when the final estimator is B ($+1.35\%$), using the same AA denominator. The final-weight replacement gives $+1.39\%$ at A's configuration and $+6.17\%$ at B's. These are

panel averages: no individual query in this group reverses its selection-effect sign. Across all 176 queries, 37 change configuration and only one shows an A-harmful, B-beneficial selection change, at case 0043/LocalNO/H10 with location jitter. Thus changes in effect magnitude across queries must be distinguished from a within-query reversal. Neither result identifies reference switching or recurrent feedback as the cause of full-trajectory differences.

The controlled table contains both local-weight failures and additional losses from configuration reselection. These effects are distinguished from reference switching and feedback, which are included only in independent recurrence.

G.2 Inferential Scope of the Mechanism Chain

The finite question is whether a beneficial final-weight replacement transfers to selection and recurrence. Equation (11) connects the first two decisions at a fixed donor state. The independent recurrence tests the resulting program after releasing that state; it is not the remainder of a telescoping causal decomposition across panels.

Table 24: What is isolated by each piece of evidence. "Controlled" refers to an intervention in the implemented reconstruction program on the evaluated data, not a physical intervention in the flow.

Evidence	Held fixed / changed	Interpretation
Final-weight replay	History, reference, configuration, observations fixed; replace A by B consistently in the solve.	Controlled conditional total effect of final location handling. It does not attribute effects to probability accuracy alone.
Conditional stage factorial	Same donor history/reference and candidate opportunity; change generator used for configuration selection and final fitting separately.	Controlled configuration effect at final A or B and their interaction. Reference selection and recurrent feedback are excluded.
Production response	Same fixed query; recover raw fields and apply the actual closure/cap.	Exact field/risk reconstruction; paired postprocessing comparisons isolate that operation. Anchor-dependent four-term allocations are explanatory accounting, not four independent causal effects.
Complete recurrence	Same observations and initial information; each frozen algorithm selects references/configurations and feeds back its own outputs.	Paired total algorithm effects and temporal stability. Reference switching and feedback are not separately identified.
Position scores / field errors	Diagnostic true locations or dense target available only to evaluation.	Location-likelihood and field-error associations do not establish mediation. Counterfactual scoring locates a mismatch but does not yield a deployable acceptance rule.

The production replay contains both positive and negative fixed-weight effects, including only 10/24 cylinder queries improving despite a favorable cylinder mean. The stage factorial reports 176 queries, whereas the acceptance audit reports 40 parent queries on different MHD/isotropic objects. They must not be pooled into a common sample or used to estimate the fraction of rollout damage caused by a given path. In the acceptance panel, all four half-observation candidates are harmful in each of 20 harmful decisions; this rules out final refitting as the origin of the harmful sign for those decisions. The 12 location-score mismatches and eight sensor-to-field mismatches remain target-assisted classifications. Sampling variability, selection optimism and spatial heterogeneity have not been uniquely separated.

The completed chain supports conditional final-stage benefits and identifies mechanisms by which they fail to transfer. A universal observation-only benefit predictor, a physical-field taxonomy, and a complete reference-versus-feedback mediation analysis are outside this claim.

G.3 A Physical Quantity Along the Correction Pipeline

For the archived uncertain-location queries of case0048 and RBC, we evaluate $E_c(f) = \frac{1}{2}\langle f_c^2 \rangle_{\mathrm{fluid}}$ and the relative error of total energy $\sum_c E_c$. The uniform-cell volume factor cancels in this relative error. We use all 16 eligible queries from the existing configuration-switch diagnostic, without ranking their energy benefit. They retain AA's history, reference and configuration. Each physical unit is averaged first, followed by equal weighting of units within each panel/backbone.

Table 25: Relative total kinetic-energy error (percent) before and after closure. The cap is inactive for these queries, so final values equal the closed values.

Panel	Backbone	A raw	B raw	A closed	B closed
0048	FNO	0.716	0.840	0.694	0.760
0048	LocalNO	0.149	0.222	0.142	0.209
RBC	FNO	21.026	21.262	21.550	21.715
RBC	LocalNO	9.489	9.382	9.614	9.476

The diagnostic does not show a uniform energy advantage from B. Field NRMSE and energy measure different errors, and closure can change their rankings differently. The fixed-state sample measures this local response; it is separate from the independent-rollout comparison.

Complete independent recurrence on 0048/FNO. The four-arm reproduction also retains final fields at all 160 queries. Averaged over H1–20 and the eight conditions, NRMSE is 0.03689 (AA), 0.03658 (cross-fit only), 0.03537 (calibration only), 0.03500 (AB). The corresponding total kinetic-energy errors are 0.252%, 0.268%, 0.264%, 0.271%. Thus lower field error in this rollout does not imply lower energy error. For a field change d, the energy change is $\langle f, d\rangle + \frac{1}{2}\langle d, d\rangle$, whereas squared reconstruction-error change depends on $\langle f - f^\star, d\rangle$. The two rankings therefore need not agree. These fields reproduce the original observation batch, rather than adding an independent data replicate.

H Repeated Observation Batches

H.1 Repeated gains and their scope

The development and confirmation batches use the same three cylinder cases with different observation randomness. AB improves mean NRMSE over AA by 2.36%/2.86% with FNO/LocalNO in the development batch and 5.25%/3.40% in confirmation. Its corresponding gains over BB are 3.25%/1.88% and 1.51%/0.71%. Each batch favors AB over AA in all three case means; against BB, the count changes from three to two. These are repeated measurements of exposed physical cases, not six independent physical samples.

The complete condition breakdown is less uniform. No nondegenerate observation condition improves over both AA and BB in every batch–backbone combination. For example, jitter with a missing component favors AB with FNO in both batches, but its confirmation mean with LocalNO is worse than AA and BB. Accurate-location conditions yield common outputs. The repeated panel advantage is confined to these original-window batches. The later-window result in Appendix P favors BB in five of six groups, so these repetitions do not establish a reliable stage-assignment rule.

Table 26: Failure frequency after averaging repetitions within physical units. Entries count unit–condition means with more than 5% higher NRMSE for AB. Accurate-location conditions remain in the denominators. Repeated observations and horizons are not independent units.

Panel	Backbone	Conditions	vs. AA	vs. BB	vs. GP
Cylinder	FNO	24	0	1	1
Cylinder	LocalNO	24	1	2	1
Magnetic	FNO	60	14	15	2
Magnetic	LocalNO	60	8	10	4
PDEArena	FNO	50	0	0	0
PDEArena	LocalNO	50	0	0	2
RBC	FNO	20	0	0	5
RBC	LocalNO	20	0	0	1
Shear	FNO	20	0	1	0
Shear	LocalNO	20	0	0	0

The two-dimensional physical units are three cylinder cases, five RBC trajectories, four shear trajectories, ten PDEArena generator files and twelve liquid-metal magnetic configurations. The three-dimensional panels add six MHD and two isotropic spatial objects; their overlapping windows are summarized separately in Table 18. Shear observation seeds and PDEArena's four trajectories per file are averaged within those units before counting wins or harmful conditions. Backbones are reported separately. These descriptive counts do not assert independence across units or supply an observation-only rejection criterion. The registered 5% threshold is used to disclose material losses; it is not a statistical significance threshold.

I Supplementary Conditional-Mixture Validation

I.1 Approximating a conditional mixture response

Fix a reference r, field covariance $G = K + (\kappa/\lambda)BB^\top$, positive working noise covariance R, and location distribution $q(Z)$. Define $A_Z = H_Z G H_Z^* + R$, $B_Z = G H_Z^*$, and $e_Z = y - H_Z r$. The average conditional field is $m(q) = r + \mathbb{E}_q[B_Z A_Z^{-1} e_Z]$. In this diagnostic, R is fixed at the original location prior; it is distinct from the production covariance κC_q. Also, $\bar{A} = \mathbb{E}_q[A_Z]$ excludes the between-location mean variance $\mathrm{Var}_q(H_Z r)$.

Write bars for expectations, deltas for deviations, $\bar{w} = \bar{A}^{-1}\bar{e}$, and

$$b_Z = \Delta e_Z - \Delta A_Z \bar{w}, \quad v_Z = \bar{A}^{-1} b_Z, \quad T_Z = \Delta B_Z - \bar{B}\bar{A}^{-1}\Delta A_Z. \tag{15}$$

The mean-matrix field is $f_0 = r + \bar{B}\bar{w}$; its second-order correction is $f_2 = f_0 + \mathbb{E}_q[T_Z v_Z]$.

Response identity. For positive-definite A_Z and finite expectations,

$$m(q) - f_2 = \mathbb{E}_q\big[T_Z(A_Z^{-1} - \bar{A}^{-1})b_Z\big]\,. \tag{16}$$

If $\bar{A} = LL^\top$, $E_Z = L^{-1}\Delta A_Z L^{-\top}$ and $\beta_Z = L^{-1} b_Z$, the integrand equals

$$-T_Z L^{-\top}(I + E_Z)^{-1} E_Z \beta_Z. \tag{17}$$

Appendix U gives the proof and a local error bound. These standard inverse identities identify three quantities to test: the relative matrix perturbation, the directions excited by the residual, and their propagation into the field. A large eigenvalue of E_Z alone does not determine the field error. Conversely, eigenvalues approaching -1 can amplify a small excited component. Accurate locations collapse the distribution to one operator and make the response difference zero.

This analysis concerns approximation of a conditional field, before the shared closure and cap. It does not predict whether that field is closer to the physical target. Even an exact working-model response can increase reconstruction error when the field model or location distribution is misspecified.

Table 27: Frozen prediction of the raw field response $m(q)-f_0$. Error is $\|f_2-m(q)\|^2/\|m(q)-f_0\|^2$ on the fluid domain.

Family	Passing instances	Median error	Median cosine
Cylinder, new observations	39/48	0.0313	0.9858
RBC, reserved trajectories	25/30	0.0179	0.9913

I.2 A FROZEN RESPONSE PREDICTION ON NEW CONDITIONS

We freeze the second-order response before evaluating new cylinder measurements on cases 0019/0048 and two reserved RBC trajectories, each with both backbones. All conditions at H1/H10 on AA histories are included. Of 96 queries, 48 have accurate locations and nine retain the original no-correction choice. The remaining 39 queries provide 78 A/B weight instances. The predictor and the average conditional reference use independent random streams, each with eight batches of 1,024 location samples. The reference uses a zero-mean first-order control variate; Taylor control variates have established precedent (Chen et al., 2018).

The registered criterion requires squared response error at most 0.1 of the reference response energy and direction cosine at least 0.9, after a sampling-resolution check. Both families exceed the registered 75% instance pass rate (Table 27). The result is uneven across physical units: 0019 passes 20/20, 0048 passes 19/28, and the two RBC trajectories pass 10/14 and 15/16. These are repeated weight/query instances, not independent physical samples.

The largest RBC failure occurs with all components observed and a weak background penalty. Subsequent fixed-query interventions show that reducing the algebraic perturbation restores the local approximation; increasing the penalty also reduces its error. Spectral projections explain a contrasting successful query: large perturbation directions exist but receive little residual excitation. These are post-hoc explanations of the frozen test, detailed in Appendix V.

Field-response accuracy and reconstruction benefit remain separate. Replacing the production B solve by the average conditional field improves only 2/15 RBC queries after common postprocessing, increasing mean NMSE by 0.00151. We therefore retain the original AB solver.

Connecting the analysis to production outputs. An additional post-hoc replay verifies Eqs. (7)–(10) on the same 96-query archive: nine no-correction queries bypass the solve, and the remaining 87 include all 39 uncertain-location queries. The final-field identity has maximum absolute discrepancy 1.1×10^{-13}; its maximum relative squared discrepancy on uncertain queries is 2.6×10^{-22}. These checks establish implementation correspondence, not new prospective prediction accuracy. Closure reduces both A's and B's absolute error in all 39 uncertain queries, yet reverses their ranking in two RBC queries. The cap changes four RBC queries, increasing both errors while favoring B relative to A. A relative method gain therefore need not mean that the intervened processing step improves either field. Appendix X gives examples and the distinction from the mixture-response test.

J ACCEPTANCE REPLAY AND SPATIAL EVALUATION

The four diagnostic tasks cover MHD and isotropic fields with FNO and LocalNO, ten parent queries per task. We replay the original and an alternative score's selected configurations using archived histories and observations. We check observation fingerprints, held scores and final candidate risks against the archives. The original score makes 7/10 harmful decisions for each MHD backbone and 3/10 for each isotropic backbone. Each harmful decision has four harmful half-observation fits. Full fitting reduces damage in 14 of the 20 decisions, and changing the final weights further reduces it in 12; these later changes do not originate the acceptance error. The backbones share physical fields, so these counts are decisions rather than independent physical samples.

For the 12 location-score mismatches, both noisy and noiseless true-position held errors increase. For the eight sensor-to-field mismatches, held errors improve but the dense visible-component error increases. All 20 therefore have visible-component damage; missing components are not its only location. The replay identifies these mismatches, but does not uniquely separate sampling variability, candidate-selection optimism and spatial heterogeneity.

A separate target-assisted check samples new, uniformly distributed diagnostic positions outside the original 64 true sensor sites, using noiseless values of the visible components. The 128 repeated samples share positions across candidates and folds, with nested budgets. In the eight sensor-to-field mismatches, the mean fraction detecting harm rises from 74.80% with 32 sites to 97.07% with 2048. For the 12 genuinely beneficial decisions, the corresponding beneficial-sign retention rises from 91.21% to 100%. These are repeated diagnostics of fixed candidates with additional information, not independent physical validation or achievable rejection rates under the original budget.

K Numerical Validation of the Production Response

We tested whether an observation-based stopping rule can control the approximation error of the implemented, postprocessed field response. The rule was frozen before evaluating six tasks: cylinder case 0043 and RBC trajectories 36 and 37, each with FNO and LocalNO references. These objects were excluded from the preceding numerical repair, although they had appeared elsewhere in the project. Reference selection, solver configurations and histories remained fixed.

The original second-order approximation met the registered relative squared error and cosine criteria in 24 of 32 resolvable responses. Residual-certified conjugate gradients met both criteria in all 32, requiring 2–16 iterations. The criteria were relative squared error at most 0.1 and cosine at least 0.9. Stopping used available observations and the working operators; the exact comparison solve and target field were excluded from the stopping decision. Six of the eight second-order failures had perturbation spectral radius below one: convergence of the expansion did not ensure adequate second-order precision.

The original implementation incorrectly certified 32 numerically unresolved responses. A separate same-query replay added the resolution guard, withheld all 32 certifications and preserved decisions for the resolvable responses. This repair adds no independent samples; the implemented bound remains a floating-point evaluation of an exact-arithmetic result.

To test direction dependence, we fixed the operators, weights, reference and original residual norm and varied the right-hand side. A preregistered next-order field score selected low- and high-sensitivity directions; all 64 signed pairs had the predicted error ordering before and after post-processing. Excited eigenvalue magnitude also ranks these pairs correctly. A separate post-hoc check therefore draws 16 equal-norm random directions independently of the scores in each of 12 active cylinder/RBC queries. For final approximation error, mean query-level rank correlation is 0.910/0.950 for the field-propagation score, versus 0.618/0.882 for the white-space remainder norm (Appendix L). This supports the role of field propagation on two exposed contexts; the artificial directions are not independent physical validation samples.

The procedure controls response precision, not reconstruction benefit. Its preparation-inclusive mean cost was 0.527 seconds versus 0.277 seconds for direct solution on the cylinder queries, and 0.524 versus 0.308 seconds on RBC. These timings exclude upstream reference prediction and selection. We therefore make no computational speed claim for the certified response calculation; nor does its precision imply that the predicted update reduces target-field error.

L Random-Direction Comparison

The initial signed-pair intervention selected eigen-directions using the next-order field score. Although it tests directional sensitivity, an excited-eigenvalue score orders the same 64 pairs correctly. We therefore supplement it with an exposed-context comparison whose directions are independent of all compared scores.

We retain the AA reference, configuration, history and A/B working systems for cylinder case 0043 and RBC trajectory 36 with FNO, at H1 and H10. In each active query, 16 standard Gaussian original-coordinate right-hand sides are normalized to the original residual norm, with fixed seed 924731. There are eight active cylinder queries and four RBC queries; the remaining 12 evaluated queries have zero response residual and no direction rows. The artificial directions never enter the recurrent history.

For each right-hand side, write $g = L^{-1}z$ and $E = M - I$ as in the main text. We compare $\|PD_BL^{-\top}E^2g\|_{\text{fluid}}$ with three simpler scores: $\|E^2g\|$, $\|g\|$ and $\|Eg\|$. The target is the norm of the difference between the second-order and direct field solutions after both closure and capping. All scores are computed without that direct solution. Table 28 averages Spearman correlations over queries within each context.

Table 28: Random-direction ranking of final approximation error. Parentheses give mean pairwise concordance, assigning half credit to score ties on distinct target errors. Queries and artificial directions are repeated interventions within two exposed physical contexts, not independent physical samples.

Context	Field propagation	$\|E^2g\|$	$\|g\|$	$\|Eg\|$
Cylinder	.910 (.903)	.618 (.748)	$-.108$ (.465)	.523 (.706)
RBC	.950 (.954)	.882 (.869)	.623 (.746)	.828 (.835)

The field score improves mean ordering in both contexts, but is not exact: individual-query correlations range from .703 to .991 on cylinder and .803 to 1.000 on RBC. It omits the inverse factor $(I+E)^{-1}$ and nonlinear cap response. The certificate controls response precision separately; neither this ranking nor that certificate predicts the sign of target-field risk improvement.

M NIGP-Inspired Vector Adaptation

This experiment implements the local input-noise linearization of McHutchon & Rasmussen (2011) in the present vector-field setting. It is an adaptation, not a run of official NIGP code. Unlike the common-grid GP moment comparator, it optimizes continuous hyperparameters by marginal likelihood and does not infer auxiliary location responsibilities.

The prior is a vector residual GP around each predictive reference. Its covariance combines a component-diagonal Gaussian kernel with a cross-component covariance from the shared training background bank,

$$G_{cc'}(x, x') = ak_\ell(x, x')\delta_{cc'} + \tau \sum_{j=1}^{32} u_{jc}(x)u_{jc'}(x'). \tag{18}$$

The basis directions are normalized by their training-bank field RMS and by $\sqrt{32}$; a fitted scale τ controls their contribution. This supplies cross-component information when one field component is unobserved. MHD uses the same periodic geometry, with a chordal Gaussian kernel. Isotropic evaluation retains its open-domain geometry and interior scoring region.

At a reported sensor location, the supplied discrete location prior gives covariance Σ_m. If J_m is the Jacobian of the reference plus a preliminary GP posterior mean, input uncertainty contributes the full visible-component block $J_m\Sigma_m J_m^\top$ to observation noise. Components at one sensor share the same displacement; replacing this block by independent scalar variances would discard that dependence. Three bounded variance multipliers, kernel length and amplitude, output noise and background scale are optimized through the full preliminary-mean gradient calculation. We use two L-BFGS-B starts (lengths 2 and 8) and a fixed 60-iteration budget. Finite budget-limited solutions remain in the experiment with convergence flags.

Residuals are normalized by the proposal residual RMS and predictions are rescaled before closure. Bounds are $\ell \in [0.5, 32]$, $\log a \in [-8, 8]$, log output-noise variance in $[-14, 3]$, $\log \tau \in [-12, 8]$, and log input-variance multipliers in $[-4, 4]$. The two starts share amplitude 1, log noise -4, log background scale -2 and unit input multipliers. Reference and background gradients use grid finite differences, with periodic differences for MHD; kernel derivatives are analytic. A 10^{-8} diagonal jitter is included in the normalized observation system.

The adaptation receives the same observations, candidate references, sensor partitions, training background, closure and cap as the archived methods. In each of four proposal/held panels it fits a GP on the proposal sensors, compares its field with the uncorrected reference using the common fixed-energy held deviance, then selects a reference and refits on all observations. Each method maintains its own recurrent state. This gives equal available observations and reference information, but not

an identical candidate family or scoring hierarchy: AB selects its grid setting with residual-scale deviance before using the fixed-energy score across references; the adaptation uses the latter score to select its two alternatives as well. The comparison therefore evaluates complete procedures rather than isolating input-noise treatment.

Before test recurrence, eight training-query checks passed, covering two datasets, H1/H2, exact locations and uncertain locations with a missing component. Checks include objective gradients, positive-semidefinite location-noise blocks, the exact-location GP limit, prediction/Gram consistency and divergence preservation. The first four test tasks were fixed as the first spatial object, first window and smallest model seed in each dataset, with both backbone families. After reading those results, we extended the unchanged method to the remaining six spatial objects under the same mechanical window/seed rule. This timing is disclosed: neither stage is an unexposed physical validation.

All 16 tasks completed, adding 760 recurrent rows. Row receipts and observation fingerprints were checked against the archived comparisons. Across 18,574 optimizer starts, 17,803 reported convergence; the remaining finite solutions were retained under the registered budget policy. This count covers both starts, all proposal panels and final refits, not independent queries.

Table 29: Paired first-window comparison, equal weighting of spatial objects within each dataset and backbone. Six MHD objects use H1–6; two isotropic objects use H1–20. One model seed per backbone. Lower NRMSE is better. NIGP+ denotes the shared-background adaptation, not official code. These rows are not the multi-window averages of Table 18.

Panel	Backbone	Condition	AA	AB	BB	GP moments	NIGP+
MHD	FNO	Exact	0.24050	0.24050	0.24050	0.24050	0.23737
MHD	FNO	Exact/missing	0.23872	0.23872	0.23872	0.23872	0.23574
MHD	FNO	Jitter	0.24278	0.23762	0.23954	0.24483	0.23449
MHD	FNO	Jitter/missing	0.24627	0.23848	0.24702	0.24831	0.23468
MHD	FNO	Jitter/noise	0.24013	0.23653	0.23775	0.24171	0.23471
MHD	LocalNO	Exact	0.24046	0.24046	0.24046	0.24046	0.23733
MHD	LocalNO	Exact/missing	0.23872	0.23872	0.23872	0.23872	0.23574
MHD	LocalNO	Jitter	0.24278	0.23762	0.23954	0.24483	0.23449
MHD	LocalNO	Jitter/missing	0.24627	0.23848	0.24702	0.24831	0.23468
MHD	LocalNO	Jitter/noise	0.24013	0.23653	0.23775	0.24171	0.23471
Isotropic	FNO	Exact	0.18577	0.18577	0.18577	0.18577	0.18952
Isotropic	FNO	Exact/missing	0.18585	0.18585	0.18585	0.18585	0.18724
Isotropic	FNO	Jitter	0.19499	0.19732	0.19764	0.19362	0.19877
Isotropic	FNO	Jitter/missing	0.19907	0.19831	0.19794	0.19557	0.19948
Isotropic	FNO	Jitter/noise	0.19565	0.19685	0.19783	0.19284	0.20017
Isotropic	LocalNO	Exact	0.18807	0.18807	0.18807	0.18807	0.18986
Isotropic	LocalNO	Exact/missing	0.18614	0.18614	0.18614	0.18614	0.18796
Isotropic	LocalNO	Jitter	0.19683	0.19810	0.19875	0.19316	0.19890
Isotropic	LocalNO	Jitter/missing	0.19574	0.19848	0.19810	0.19413	0.19952
Isotropic	LocalNO	Jitter/noise	0.19568	0.19761	0.19845	0.19355	0.19921

On MHD, the adaptation has lower mean NRMSE than AB for every spatial object and backbone after averaging the five conditions. AB's condition-mean excess in the three uncertain-location settings is 0.78–1.62%. However, the adaptation selects the uncorrected reference on 346 of 360 MHD steps, including 214 of 216 uncertain-location steps. Its advantage therefore cannot be credited to successful NIGP corrections without a further matched intervention. The original grid procedures can select an update even when this different selection procedure retains the reference.

On isotropic turbulence, AB has lower five-condition means on both objects with both backbones. Its uncertain-location mean improvements over the adaptation range from 0.40% to 1.66%; the GP moment comparator nevertheless has lower means than AB in these same conditions. Thus the additional route does not yield a universal winner or establish general superiority of AB over uncertainty-aware GP procedures. The correction-versus-reference decision, covariance model and hyperparameter selection differ jointly here. A fixed-donor replay would be needed to assign these complete-procedure differences to one of them.

N PANEL-SPECIFIC RECURRENT RESULTS

The complete registered panel table is retained below, including all negative comparisons. These rows use the original frozen references; subsequent reference-strength interventions are reported separately in Appendix N.1.

Table 30: Common-base and internal comparisons: complete recurrent relative NRMSE reduction of AB. Positive values favor AB. Cylinder, RBC and shear use H1–20; PDEArena uses H1–12 and magnetic fields H1–19. Units and conditions are specified in the evaluation protocol.

Panel	Backbone	vs. AA internal	vs. BB internal	vs. GP moments common program
Cylinder	FNO	5.25%	1.51%	2.28%
Cylinder	LocalNO	3.40%	0.71%	1.22%
RBC	FNO	−0.56%	−0.28%	−2.54%
RBC	LocalNO	−0.54%	−0.07%	−0.71%
Shear	FNO	1.20%	−0.72%	4.89%
Shear	LocalNO	2.23%	0.02%	6.30%
PDEArena	FNO	−0.06%	0.11%	−0.34%
PDEArena	LocalNO	0.09%	−0.05%	−0.76%
Liquid metal	FNO	−2.48%	−1.55%	6.33%
Liquid metal	LocalNO	0.47%	0.76%	8.84%

On the cylinder panel, AB improves over AA in all three cases for both backbones, and over BB and GP moments in two of the three. Case 0048 improves over AA by 5.10% and 4.05%, but is worse than BB by 1.21% and 0.60% and worse than GP moments by 0.62% and 2.70%. Figure 6 reports the individual cylinder and RBC effects.

The RBC transfer result is negative for the added design: every trajectory in each backbone has higher mean error under AB than under AA. The increase is small but consistently directed. This cohort favors retaining AA over AB.

The shear panel from The Well (Ohana et al., 2024) contains four trajectories, two backbone families and two observation seeds, each evaluated over five conditions and H1–20. References are trained at Reynolds number 10^5 and evaluated at 10^4, using a shared periodic-domain adapter and a training-increment POD32 bank. Each backbone has one training seed. AB improves the task mean over GP in all 16 combinations and over AA in 14, but over BB in only seven. Its mean improvement over AA and GP remains positive after omitting any one physical trajectory. Relative to BB, FNO deteriorates in both observation seeds, while LocalNO changes sign between seeds and is nearly tied overall. The periodic operator, kernel and location neighborhoods are common to the compared correction methods; accurate-location controls coincide numerically.

PDEArena provides little additional gain from AB over AA: their means differ by less than 0.1% for each backbone. AB is worse than GP moments by 0.34% and 0.76%, and this ordering remains after removing any one of the ten source files. The approximately 23% improvement over the projected neural reference is therefore a property of the reconstruction system, not evidence for the added AB design.

On liquid-metal magnetic fields, AB improves over GP moments in 11 of 12 FNO configurations and all 12 LocalNO configurations. Its comparison with the auxiliary alternatives differs: FNO is worse than AA and BB on average, whereas LocalNO improves by 0.47% and 0.76%. The gain over GP moments and the gain from separating stages are distinct comparisons.

N.1 REFERENCE STRENGTH AND A BOUNDED AMPLITUDE REPAIR

Same-checkpoint reference controls. Six cylinder and ten RBC tasks replay the exact checkpoints, sensor streams and initial fields used by the frozen AB comparison. Native and physically closed references each retain their own recurrent history. Every observation fingerprint matches the AB archive; the original AB first step is reproduced before running the controls. The closed refer-

ence uses the same increment projection as AB. These are task-adapted predictors at the registered checkpoints, rather than claims about fully optimized FNO or LocalNO architectures.

Table 31: Full H1–20 reference controls on the current cylinder and RBC panels. Lower NRMSE is better.

Panel	Backbone	Native	Closed feedback	Frozen AB
Cylinder	FNO	0.150952	0.148972	0.021115
Cylinder	LocalNO	0.147540	0.145038	0.020513
RBC	FNO	0.740832	0.734551	0.391273
RBC	LocalNO	0.374085	0.353105	0.289452

AB improves these frozen references on both panels. RBC nevertheless favors AA over AB: complete-system gain and the added location-estimator gain are different comparisons.

PDEArena training sensitivity. The original small references were selected after 1,600 single-step training updates. New controls use 4,800 updates with checkpoint selection on the same separate validation set. Small FNO and LocalNO retain their original architectures; a larger FNO increases hidden width and spectral bandwidth to approximately the RecFNO parameter budget. Recurrent updates supervise all twelve steps with detached feedback, so update counts do not imply equal computational or frame-exposure budgets. All three recurrent controls select step 4,800; this does not establish convergence.

On all forty test trajectories, unclosed reference NRMSE is 0.52424 for small FNO, 0.48783 for LocalNO and 0.47246 for larger FNO. The paired RecFNO adaptation gives 0.48638 and FLRONet 0.51854. The original frozen AB errors are 0.66891/0.62124. These results show that the initial external deficit is sensitive to reference training and capacity; the new reference scores are not new full-dataset AB scores.

Four objects (indices 0, 3, 6 and 9, trajectory zero) were specified before these new reference results. Replacing only the reference checkpoint while retaining the original calibration, basis, candidate grid and coefficient process reduces AB error from 0.62744 to 0.49565 for small FNO, from 0.58705 to 0.46758 for LocalNO, and to 0.45309 for larger FNO. Each arm has its own H1–12 history. Yet all three now slightly worsen their own closed reference (Table 32). On the same four objects RecFNO has native/postclosed errors 0.46447/0.45945. This targeted intervention does not establish full-dataset or seed-repeated superiority of the revised system.

Where harm enters. First-step replay reproduces all twelve intervention tasks, fixing the initial history. For larger FNO with accurate locations and a missing second component, the mean normalized risk cross-term is -0.003386 and update energy is 0.006220, giving a risk increase of 0.002834. The selected reference equals the closed predictor in these queries. A beneficial average direction can therefore be outweighed by update magnitude without location uncertainty or recurrent accumulation. Some other conditions have a positive cross-term. The implemented cap scales with field amplitude, not remaining prediction error, and held-sensor selection does not directly constrain the missing component's full-field loss.

Calibration-only amplitude adjustment. A bounded follow-up replaces the final field by $r + \gamma(f - r)$, retaining the selected reference r, configuration, location weights and state-update rule. Four validation trajectories (indices 0, 25, 50 and 75) determine γ from $\{0, 0.125, 0.25, 0.5, 0.75, 1\}$ by mean NRMSE along the original AB donor history. We freeze two variants: one global coefficient and coefficients grouped by full versus missing components. These validation trajectories are separate from test but were also used for checkpoint selection. Both variants are evaluated without selecting between them on test.

Table 32: PDEArena amplitude sensitivity on four exposed objects, full H1–12. Original AB uses the new reference with the old downstream configuration. Lower NRMSE is better.

Reference	Closed reference	Original AB	Global shrinkage	Mask shrinkage
Small FNO	0.494095	0.495647	0.494897	0.491838
LocalNO	0.453850	0.467576	0.466160	0.459705
Larger FNO	0.442255	0.453092	0.447975	0.445175

All global coefficients select 0.75. Full-component coefficients select 1; missing-component coefficients select 0.5 for small FNO and 0.25 for LocalNO and larger FNO. Each test arm runs independently, and original-reference/AB outputs reproduce the archived trajectories. Both adjustments improve original AB in all three aggregate means. Mask shrinkage slightly improves the small-FNO reference but still harms the LocalNO and larger-FNO references. This is a post-hoc repair on previously exposed objects, not a replacement of frozen AB or a fresh generalization result.

Liquid-metal reference sensitivity. We also strengthen the liquid-metal FNO reference while retaining its original architecture. Training uses 4,800 detached recurrent updates over H1–19, with checkpoint selection on the separate calibration split. CAL selects step 4,800. Four predetermined objects (indices 0, 3, 6 and 9) then replace only the reference checkpoint; the auxiliary calibration, basis, candidate grid and coefficient process remain fixed. Each reference and AB arm generates its own history, and all 760 new evaluation rows match the original observation fingerprints.

Table 33: Liquid-metal reference sensitivity on four objects, H1–19. Original and strengthened references use the same FNO architecture. Reference columns include physical closure and independent feedback. Lower NRMSE is better.

	Original reference		Strengthened reference	
Condition	Reference	AB	Reference	AB
Clean	0.449995	0.069992	0.138452	0.063518
Location jitter	0.459114	0.112056	0.139572	0.086385
Jitter + noise	0.453488	0.112737	0.140776	0.090766
Missing second component	0.446235	0.157026	0.144806	0.091045
Jitter + missing component	0.459629	0.204056	0.142516	0.116970
Mean	0.453692	0.131174	0.141224	0.089737

With the strengthened reference, AB reduces aggregate NRMSE by 36.46%, and improves the mean for each of the four objects. Thus stronger references do not invariably remove the value of correction: the liquid-metal intervention differs from PDEArena. These four-object sensitivity results retain the original downstream estimator and do not replace the full-panel frozen comparison.

O Fixed-Reference Component Interventions

An earlier controlled evaluation compares six location rules on the three cylinder cases with two backbones: prior weights, reference-derived weights, auxiliary weights, auxiliary cross-fitting alone, likelihood calibration alone, and their combination. The reference, history and full kernel solver are shared. Each model contributes eight conditions at H1 and H10, for 96 queries. Table 34 isolates changes at the donor's fixed configuration; these are development interventions, rather than independent recurrent trajectories.

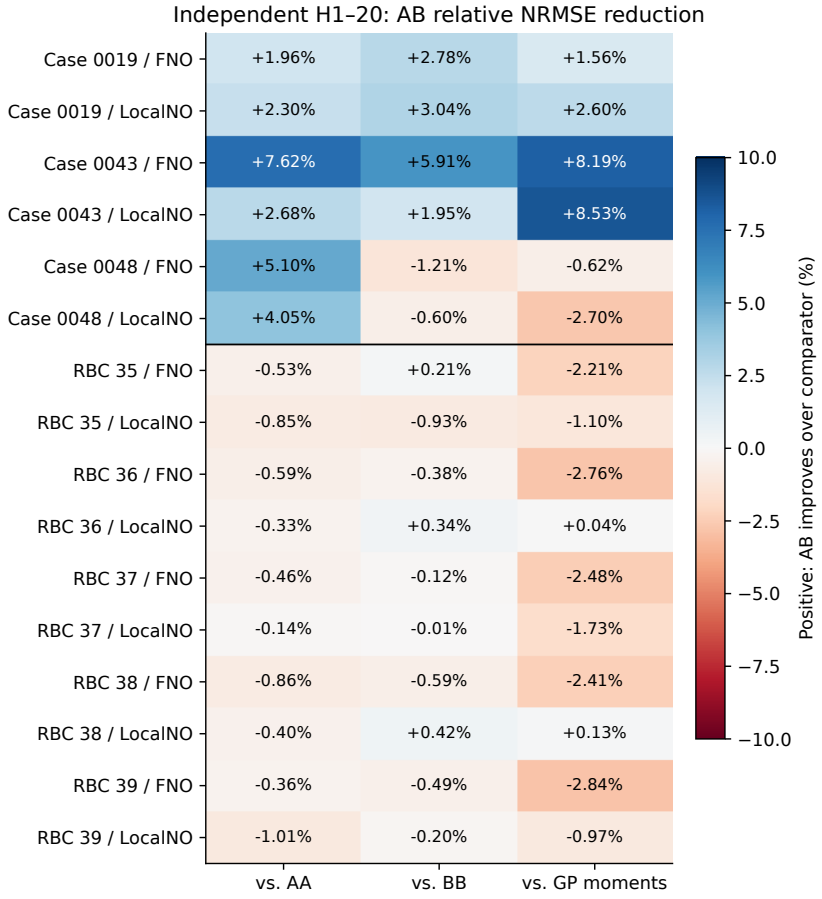

Figure 6: Physical-unit effects for the frozen independent recurrence. The horizontal line separates cylinder cases from new RBC trajectories. Each physical unit is evaluated with both backbone families.

Table 34: Relative NRMSE reduction over original auxiliary weights at a shared reference and fixed configuration. Positive values favor the modification.

Case	Backbone	Cross-fit only	Calibration only	Both
0019	FNO	0.03%	0.29%	0.47%
0019	LocalNO	0.03%	0.31%	0.39%
0043	FNO	0.05%	0.06%	0.82%
0043	LocalNO	3.48%	−0.53%	4.51%
0048	FNO	0.68%	−0.56%	2.74%
0048	LocalNO	1.87%	−0.44%	4.58%

Cross-fitting improves all six conditional means, whereas calibration of the original auxiliary estimator helps only three. The combination improves all six means over either single change. Allowing configuration reselection at the same reference changes this ordering: the combination becomes worse than calibration alone for 0019/LocalNO and 0043/FNO, and worse than cross-fitting alone for 0048/LocalNO. Thus a favorable component comparison at a fixed configuration does not settle its role in the complete selection procedure.

Table 35: Later-window relative NRMSE reduction of frozen AB. Physical cases receive equal weight. Positive values favor AB.

Backbone	vs. AA	vs. BB	vs. GP moments
FNO	1.79%	−1.54%	0.23%
LocalNO	0.70%	−0.70%	0.38%

Table 36: Median H1 batch time in seconds, with the range of three repetitions. All methods use the same GPU.

Context	AA	AB	BB	GP moments
Cylinder/FNO	43.00	46.81	43.06	39.55
Range	42.91–43.37	46.71–46.92	43.05–43.24	39.52–39.73
RBC/FNO	15.32	17.22	15.93	13.97
Range	15.31–15.37	17.13–17.23	15.92–15.96	13.97–13.98

P Later-window Confirmation and Adaptive Development Check

We retain the frozen methods and evaluate frames 744–765 of the same three cylinder cases, using two initial fields and H1–20, eight conditions and a new observation seed. The six tasks contain 960 paired queries with independent method histories. This window does not overlap the original frames 244–265, but the physical cases have development exposure. Time and observation seed change together. AB improves over AA in all six case–backbone groups, whereas BB outperforms AB in five. Both backbones fail the preregistered criterion requiring positive mean improvement over both AA and BB and nonnegative effects in at least two of three cases. Case 0048 also favors GP moments over AB with both backbones. This result limits the original-window finding: replacing the final estimator retains an aggregate benefit, but retaining A for selection is not reliably preferable to using B in both stages. We do not change the AB definition or reinterpret BB as a newly confirmed method.

A subsequent development study tested observation- and candidate-based selection among AA, AB and BB on common AA histories. Ridge and shallow tree predictors were evaluated with entire physical objects held out, across 12 liquid-metal configurations and the three cylinder cases. Neither met the prespecified cross-object retention criteria. This one-step study does not evaluate an adaptive closed-loop method; it provides no basis for claiming that an automatic selector resolves the observed stage dependence.

Q Measured Inference Cost

We time an identical H1 workload for each method in a fresh process on one NVIDIA A800, with one warmup and three measured repetitions. Cylinder and RBC contexts use two and one PyTorch CPU threads, respectively, with the thread count fixed across methods within each context. Table 36 reports median batch time; cylinder has eight conditions and RBC has four. Each repetition reproduces the frozen reference, configuration and reconstruction metric. The separated implementation costs 8.9% and 12.4% more than AA in these workloads. It currently computes the original candidates before replacing the final weights. These measurements describe the implemented H1 workloads, not full-horizon scaling. Peak allocated GPU memory is 72.6 MiB for cylinder and 53.2 MiB for RBC for all four methods, including the resident model. Process peak resident memory ranges from 1721 to 1752 MiB and 1677 to 1707 MiB, respectively; it includes context initialization and warmup. Shared GPU activity was recorded before each task, but hardware contention cannot be excluded completely.

R Condition-Level Failures

Table 26 reports how often AB exceeds each comparator by more than 5% after averaging repetitions within physical units. On magnetic fields with FNO, this occurs in 14 of 60 unit–condition means

against AA, but only two against GP moments. The choice of comparator therefore changes both the apparent benefit and its failure rate. The individual-repetition failures below remain visible even when averaging reduces them. The complete rollout also contains material condition-level harm. For case 0048 with LocalNO, jitter and the missing v component, AB increases NRMSE by 7.80% over AA and 33.27% over GP moments. The first exceeds the prespecified 5% harm-disclosure threshold. A positive case average does not make the update safe across conditions.

In shear flow, FNO trajectory 1 with location jitter and missing v increases error over AA by 2.86% and 6.05% in the two observation seeds. The second also exceeds GP error by 0.74%, despite AB beating GP in every task mean. The largest AB–BB condition-level increase is 8.91%, for LocalNO trajectory 2 in the second seed with the same observation condition.

The additional panels also contain substantial conditional losses. On PDEArena, AB increases error over GP moments by 23.89% and 25.38% in the worst FNO and LocalNO trajectory–condition pairs, both with location jitter. In magnetic fields, the FNO configuration at magnetic Reynolds number 200 and imposed field 0.50 has 46.78% higher error than AA with jitter and a missing component; the worst LocalNO loss against GP moments is 25.30%. Mean gains cannot be used as a rule for accepting an update.

Earlier controlled experiments on a separate finite-dimensional auxiliary estimator provide complementary evidence. With field and sampled error energies matched, high-transfer out-of-space perturbations produce greater auxiliary contamination on five independently reserved cross-Rayleigh trajectories. Joint nuisance estimation reduces that contamination contrast but does not consistently improve the downstream correction. These interventions use true positions to test contamination under controlled perturbations.

S Absolute Reconstruction Errors

Table 37: Absolute mean field NRMSE for the current methods. Each panel uses its own registered observation conditions and physical units.

Panel	Backbone	AA	AB	BB	GP moments
Cylinder	FNO	0.022285	0.021115	0.021440	0.021608
Cylinder	LocalNO	0.021236	0.020513	0.020659	0.020767
RBC	FNO	0.389083	0.391273	0.390199	0.381586
RBC	LocalNO	0.287896	0.289452	0.289259	0.287424
Shear	FNO	0.068206	0.067389	0.066909	0.070854
Shear	LocalNO	0.066817	0.065327	0.065341	0.069720
PDEArena	FNO	0.668520	0.668906	0.669650	0.666670
PDEArena	LocalNO	0.621807	0.621241	0.620907	0.616576
Liquid metal	FNO	0.138383	0.141813	0.139648	0.151396
Liquid metal	LocalNO	0.129981	0.129366	0.130362	0.141907

The relative reductions in the main text are calculated from these unrounded means. Physical units receive equal weight within each panel. Cylinder uses eight observation conditions, RBC four, and the other panels five; absolute values across panels therefore describe different tasks rather than a pooled ranking.

T Complete Development Factorial

For the subsequent seed-5100 panel, the recorded development decision rule required at least 0.5% mean improvement over AA and BB in each backbone family, at least two of three cases non-negative, and at most 2% mean excess over the GP moment comparator. These were engineering retention criteria, not significance or equivalence tests.

The development evaluation used observation seed 0, 64 sensors, all eight registered conditions and H1–20 on the same three cylinder cases. Each of AA, AB, BA and BB generated its own recurrent history. Table 38 reports the relative reduction in each arm's mean NRMSE compared with AA. These observations participated in method development and are distinct from the seed-5100

confirmation panel in the main text. AB improves the case-level mean in all six case–backbone

Table 38: Development four-arm recurrence. Positive values favor the indicated arm over AA.

Case	Backbone	AB	BA	BB
0019	FNO	$+0.843\%$	-3.159%	-0.798%
0019	LocalNO	$+0.967\%$	-3.006%	-0.739%
0043	FNO	$+6.612\%$	$+1.474\%$	-0.362%
0043	LocalNO	$+5.419\%$	-9.448%	$+4.469\%$
0048	FNO	$+0.482\%$	-2.780%	-1.274%
0048	LocalNO	$+2.228\%$	-3.578%	-0.120%

combinations, whereas BA worsens five. The positive BA result for case 0043 with FNO is a counterexample to a universal prohibition on using B for selection. The two backbones share physical cases and are not independent physical samples.

The archive contains 960 recurrent rows and a separate 96-query fixed-reference intervention at H1 and H10. We checked result receipts and source hashes, recomputed the squared-error factorial identities, and verified the H1 selection agreement between BA and BB. For the 43 H1 queries where the selected reference also matches the conditional intervention, BA's field risk agrees numerically. These checks validate the relationship between the stored evaluations without rerunning or retuning the methods.

The main-text conditional table uses new observation randomness and fixes AA's reference and history. The development recurrence changes both reference selection and subsequent history. Subtracting these two tables would mix observation differences with different interventions; it would not isolate a feedback effect. A stronger claim about reference switching or history amplification requires a matched donor-history intervention.

U Response Identities and Scope

All expectations in this section hold the reference, configuration and working field model fixed. The empirical quadrature uses the product of sensor-location marginals, with a single shared location for a sensor's visible components. It is not a jointly inferred location posterior.

Proof of Eq. (16). Let $u_Z = A_Z^{-1} b_Z$. Then $A_Z^{-1} e_Z = \bar{w} + u_Z$ and $u_Z - v_Z = -\bar{A}^{-1} \Delta A_Z u_Z$. Expanding $B_Z = \bar{B} + \Delta B_Z$ and subtracting $\bar{B}\bar{w}$ gives a first-order term $\Delta B_Z \bar{w} + \bar{B} v_Z$, whose expectation is zero, plus $T_Z u_Z$. Hence $m(q) = f_0 + \mathbb{E}[T_Z u_Z]$, and subtraction of f_2 proves the identity. Substitution of $A_Z = L(I + E_Z)L^\top$ proves Eq. (17).

If $\|E_Z\|_2 \le \rho < 1$ for every supported location state, the Neumann bound yields

$$\|m(q) - f_2\|_2 \le \frac{\rho}{1-\rho} \, \mathbb{E}_q \left[\|T_Z L^{-\top}\|_2 \|\beta_Z\|_2 \right]. \tag{19}$$

The condition is sufficient, not necessary. The exact expression retains directional projections and cancellations discarded by this norm bound. In particular, a direction with a large relative eigenvalue contributes little if its residual excitation or field propagation is small. A location distribution concentrated at one state has zero deltas and zero approximation gap.

Probability accuracy is not a field-risk guarantee. In a specified Gaussian working model, let $p^*(Z \mid y)$ be the exact location posterior and $\mu^* = \mathbb{E}_{p^*}[\mu_Z]$, where $\mu_Z = \mathbb{E}[u \mid y, Z]$. For a fixed positive-semidefinite loss matrix Ω, conditional quadratic risk obeys

$$\mathcal{R}(F(q) \mid y) - \mathcal{R}(\mu^* \mid y) = \|[m(q) - \mu^*] + [F(q) - m(q)]\|_\Omega^2. \tag{20}$$

This is the standard Bayes quadratic-risk identity. Improving the posterior approximation can alter its cancellation with the solver approximation. The identity requires a loss matrix fixed given y; it is not directly a theorem about target-normalized NRMSE or a misspecified physical model.

A scalar example makes the distinction explicit. Let $a \sim N(0,1)$, $Z \in \{1,2\}$ with equal prior probabilities, and $y = Za + \epsilon$, $\epsilon \sim N(0, 0.1)$. For $y = 1$, $p^*(Z = 1 \mid y) = 0.58060176$ and $\mu^* = 0.73240429$. The mean-operator decision $F(q) = h/(h^2 + 0.1)$, $h = \mathbb{E}_q[Z]$, equals μ^* at $q_0(Z = 1) = 0.71229119$. Along the segment from q_0 to p^*, posterior KL and expected Brier excess decrease to zero, while the conditional excess quadratic risk of F increases from zero to 0.00374496. The mixture decision $m(q)$ instead approaches μ^*. This constructed working-model example establishes possibility; it is not a deployable estimator or a claim that the general decision-theoretic principle is new.

V Response Validation and Failure Audits

Frozen validation. The cylinder panel uses observation offset 7100 on existing cases 0019/0048. The RBC panel uses physical trajectories 7/8, absent from the audited source-reader and registered trajectory ranges, with frames 50–61. This audit cannot exclude unrecorded external data access. Both backbones, 64 sensors and all existing conditions are retained at H1/H10. AA generates the reference, configuration and H1–10 history; neither diagnostic output nor target field enters its recurrence.

All 48 accurate-location controls reproduce the common solve. Nine rank-zero choices remain abstentions. All 78 uncertain-location weight instances have reference-response energy exceeding nine times the estimated mean-field sampling variance and the registered numerical floor. The predictor and reference use independent random streams; A/B share random numbers within each stream. The criterion requires squared relative response error at most 0.1 and cosine at least 0.9, with at least half the eligible instances resolvable and 75% of resolvable instances passing within each physical family. These are fixed engineering criteria, not calibrated confidence statements. No outcome-dependent quadrature expansion is used.

A separately registered disconnect test. On earlier new measurements of 0019/0048, changing A to B improves both true-location NLL and Brier score but increases final NMSE in 4/32 uncertain-location queries. This occurs in 1/16 queries on reserved RBC trajectories 5/6. Four of the five events have all components observed. A prespecified masking intervention predicted that re-estimating weights after removing v would worsen B relative to A compared with freezing the full-component weights. The result has the opposite sign in all four physical units. We reject that directional hypothesis. Oracle true-position replacement reduces mean error in each unit but does not supply an observation-only repair.

Why the largest response error occurs. For RBC trajectory 7/FNO/H1 with location jitter, A and B both fail. The selected setting is rank32, ridge 10^{-4}, kernel scale 0.1 and length 2. The B mean-matrix condition number is about 27,422; the median sampled relative perturbation norm is 10.61. This indicates large variation relative to the expansion point, not a demonstrated floating-point solve failure. Exact coefficient identities agree within numerical tolerance.

Using fixed paired location samples, B has squared response error 2.278 and cosine 0.644. Along the algebraic path $(\bar{B} + t\Delta B, \bar{A} + t\Delta A, \bar{e} + t\Delta e)$, these become 0.0398/0.9985 at $t = 0.025$ and 0.00677/0.9997 at $t = 0.01$. This path tests the local expansion, not a new physical measurement process. Holding the reference, weights and all other settings fixed, increasing ridge to 0.1 reduces the error to 0.120 and raises the cosine to 0.961, still outside the original error criterion. The normalized covariance variation is dominated by the POD observation response; the kernel-to-POD RMS Frobenius ratio is 5.44×10^{-5} in this query.

Spectral projection distinguishes large perturbations from excited perturbations. In 128 sampled states, the failing query places on average 76.2% of the whitened driving energy in directions with relative eigenvalue above 1. The successful LocalNO query at the same physical trajectory and horizon places only 0.238% there, although 47/128 states have an eigenvalue above 1. The successful FNO query on trajectory 8 has 4.95% in those directions. Negative eigenvalues near -1 and field propagation also matter: the smaller FNO H10 failure is dominated by the former. These selected-query diagnostics do not establish a universal spectral rejection threshold.

Retaining the inverse response. A fixed-budget exploratory preconditioned conjugate-gradient calculation approximates $A_Z^{-1} b_Z$ instead of substituting $\bar{A}^{-1} b_Z$. At 32 iterations, the largest failure's paired-sample squared response error decreases from 2.793 to 0.0774 and cosine rises from 0.609 to 0.971. The other three diagnostic queries also approach the exact response. Field error is not monotone in the iteration count. We have not established a speed advantage over direct solves or a reduction in physical reconstruction risk, and this calculation is not included in AB.

W Why a Weakly Regularized Candidate Was Selected

We exactly replay the RBC7/FNO/H1 selector, reproducing its observation fingerprint, reference weight $\alpha = 0.5$ and candidate97. Its mean held mixture deviance is 2.57919×10^{-5}, only 0.208% lower than runner-up candidate53. Two of four panels choose it. Omitting one panel at a time changes the winner in two of four cases; the panels overlap and are not independent statistical repetitions.

Table 39: Post-hoc risk comparison at the same full reference. Dense targets and true locations are used only for diagnosis. Half-fit risks average the four proposal fits.

Candidate	Half-fit field NMSE	Full-fit field NMSE	True-location held MSE
Selected97	0.06251	0.03943	1.3522×10^{-5}
Runner-up53	0.03871	0.03405	1.5728×10^{-5}
Full-risk best89	0.03723	0.03150	1.5423×10^{-5}

The selected candidate already has larger field error in the half-fit stage. Its advantage on held sensor values persists even when evaluated at the true sensor locations. Thus the discrepancy cannot be attributed solely to mistaken position explanations or to the final 64-sensor refit. Finite sparse validation and full-field risk rank these candidates differently. Replacing proposal references by the common full reference changes the score winner to candidate53, but this diagnostic can compromise the original sensor holdout and is not proposed as a new selector.

Training-only follow-up. An exploratory check restricts an existing bounded-preference full-risk ranker to candidates whose mean score difference from the original winner is no larger than one paired-panel standard deviation. This spread is a heuristic, not a standard error or confidence interval. Two existing training objects per FNO context are alternately used for fitting and evaluation; no held-test files are read. Relative to original selection, the pooled NRMSE changes favor the constrained ranker by 4.72% for the 0048 context, disfavor it by 3.60% for 0019, and favor it by 0.41% for RBC. RBC changes sign across the two transfer directions. A training configuration-risk prior likewise fails to transfer across all contexts. Neither arm meets the predeclared advance criterion, so neither is tested in a new recurrent method. This result leaves stable full-field candidate-risk selection unresolved.

X Production-Response Verification

The replay fixes AA's reference, history, configuration, current observations and A/B location weights. It reconstructs the original raw fields, solves the equivalent dual systems in Eq. (6), and separately computes propagation, residual, Gram and working-covariance terms. Each term is passed through the common linear closure, multiplied by B's cap factor, and combined with the cap-scale difference in Eq. (9). Their sum reproduces the archived final-field difference. The signed final-risk allocation reproduces the directly evaluated NMSE difference. The no-correction branch is treated as an explicit bypass rather than assigning a fictitious positive-rank solve.

This finite-change analysis differs from the mixture-response validation. It includes the actual q-dependent working covariance and nonlinear cap; it does not average conditional location solves or predict a previously unseen outcome. Its target-dependent risk terms explain observed losses. Moreover, the choice of A as the expansion anchor changes individual terms, and propagation and Gram contributions can be large and opposing. A term-wise risk allocation is not the same estimand as changing one solver component and resolving both fields.

Table 40: A common processing step can reverse A/B rankings without improving the relative winner's absolute field quality. All values are NMSE at fixed references and configurations.

Query	Output	A	B
RBC8/LocalNO, H10, jitter	Raw	.263030	.261692
	Closed	.231498	.231975
RBC7/FNO, H10, jitter	Closed	.261344	.262464
	Capped	.278939	.278859
0048/LocalNO, H10, jitter/missing v	Mean matrix	.003031	.002522
	Mean operator, fixed noise	.007262	.003740

In the first example closure helps both fields but helps A more. In the second the cap hurts both fields but hurts A more. In the third replacing the same-sensor second moment by the mean-operator Gram worsens both fields while increasing B's relative advantage. None alone establishes a better independently recurrent replacement.

The uncertain-query averages also conceal heterogeneous effects. B improves final NMSE in 10/24 cylinder queries despite a favorable cylinder mean; removing the largest improvement reverses that mean. In RBC, B improves 9/15 queries but the mean deteriorates because larger losses outweigh the gains. At H1 all seven eligible RBC queries improve; at H10 only two of eight do. These are AA donor histories, so this contrast does not identify feedback from B's outputs as its cause.

Among queries where both A and B pass the frozen mixture-response approximation criterion, B's production field is worse in 11/19 cylinder and 6/12 RBC queries; eight and one, respectively, exceed a 1% NRMSE increase. This post-hoc cross-tabulation demonstrates the distinction between the two evaluation targets. It is not a newly registered risk-classifier test or evidence of independent physical repetitions.

X.1 The error geometry behind a closure reversal

For the common affine closure $T(f) = u_0 + P(f - u_0)$, let $a = f_A - u$, $v = f_B - f_A$, and $b = T(u) - u$. Before the cap, the A/B loss differences are exactly

$$\Delta_{\mathrm{raw}} = 2\langle a, v\rangle_\Omega + \|v\|_\Omega^2, \tag{21}$$

$$\Delta_{\mathrm{closed}} = 2\langle Pa + b, Pv\rangle_\Omega + \|Pv\|_\Omega^2. \tag{22}$$

Thus a raw gain reverses precisely when $\Delta_{\mathrm{raw}} < 0 < \Delta_{\mathrm{closed}}$. The vector v is supplied by Eq. (7), so this criterion links location-weight changes to the implemented observation operator, covariance and closure. The offset b matters when the evaluation target does not exactly satisfy the discrete closure. Even when $b = 0$, closure replaces the evaluation geometry by $P^*\Omega P$.

A useful special case explains why improving both absolute errors need not preserve their ranking. If P is an orthogonal projection in the evaluation norm and $T(u) = u$, orthogonality gives

$$\Delta_{\mathrm{closed}} = \Delta_{\mathrm{raw}} - \left(\|(I-P)(f_B - u)\|_\Omega^2 - \|(I-P)(f_A - u)\|_\Omega^2\right). \tag{23}$$

A raw improvement can therefore be concentrated in error removed from both outputs by closure; the error that survives closure can favor A. This special-case interpretation requires orthogonality in the actual evaluation norm, not merely in a full-grid norm. Equation (22) remains valid without that assumption. Active unequal caps require the full Eq. (9) instead.

Figure 2 tests the unrestricted identity on a fixed query with inactive caps: both location scores and raw field error improve, closure reverses the field ranking, and the final harm is concentrated in the unobserved component. This connects the probability-induced response to the part of field error retained by closure. The target-assisted calculation explains the observed sign; it does not provide an observation-only acceptance rule.

Y Working Covariance, Penalties and Calibration

The location weights enter a linear observation operator before the kernel solve. This appendix specifies the working covariance and normalizations used in that solve. For a fitting panel with N

visible scalar measurements, let x_{m0} be the reported location and define

$$v = \max\left\{ \frac{1}{N}\sum_{m,c} M_{mc}\big(y_{mc} - r_c(x_{m0})\big)^2, 10^{-4}\frac{1}{N}\sum_{m,c} M_{mc}y_{mc}^2, 10^{-12} \right\}. \tag{24}$$

This is a residual scale computed from the supplied reference and visible data, not a known measurement-noise variance. For location weights q, set

$$\bar{r}_m = \sum_j q_{mj} r(x_{mj}), \tag{25}$$

$$\Sigma_m(q) = \sum_j q_{mj}\big(r(x_{mj}) - \bar{r}_m\big)\big(r(x_{mj}) - \bar{r}_m\big)^\top. \tag{26}$$

The visible scalar entries of C_q are

$$(C_q)_{(m,c),(n,d)} = \delta_{mn}\delta_{cd} + \delta_{mn}\frac{\Sigma_m(q)_{cd}}{v}. \tag{27}$$

Thus the two components of one sensor share a location covariance. Missing components are omitted from the system.

The kernel regularization is $\kappa = N\xi$, with $\xi \in \{10^{-6}, 10^{-3}, 10^{-1}\}$. For background rank p, let $E_k = \sum_{m,c} M_{mc}[B^{\mathrm{raw}}_{k,c}(x_{m0})]^2$ be the sampled raw-basis energy. The coefficient penalty is $\lambda = \rho p^{-1}\sum_{k=1}^{p} E_k$, with $\rho \in \{10^{-4}, 10^{-2}, 1, 10^{-3}, 10^{-1}, 10\}$ in the listed grid order. The ranks are $8, 16, 32$ and Gaussian kernel lengths are 2 and 8 grid cells. A no-correction candidate precedes the 108 nonzero candidates. These settings and their normalization are shared across the location-weight comparisons; background energy is measured at the reported sites, independently of the inferred weights.

After the linear solve gives f, the common closure produces $f_c = f_0 + P(f - f_0)$, where f_0 is the initial complete field and P is the dataset's discrete increment projection. With $\Delta = f_c - r$, the returned field is

$$r + \min\left\{1, \frac{0.25\max(\mathrm{RMS}(r), 10^{-8})}{\mathrm{RMS}(\Delta)}\right\}\Delta, \tag{28}$$

with scale one when $\Delta = 0$. The no-correction candidate returns r directly.

Cross-fitting partitions the ordered sensors of each fitting panel by index modulo four. The auxiliary estimate excluding a sensor's block supplies that sensor's likelihood. For each visible-component mask and auxiliary estimator, the calibration collects standardized residuals at known training sensor locations. It uses their sample mean and population variance, clipping each visible-component variance to $[0.25, 16]$. Hidden components use mean zero and variance one and do not enter the likelihood. The predictive reference and history remain shared inputs to these auxiliary fits, so this construction does not remove every dependence on a sensor from the reconstruction pipeline.

Candidate selection retains the supplied location prior. Within a reference, held mixture deviance uses the proposal reference's residual scale v. Across references, the held score instead uses a common variance $\max(10^{-4}E, 10^{-12})$ and divides by E, where $E = \max(N^{-1}\sum M_{mc}y_{mc}^2, 10^{-12})$ is computed on the proposal panel. This distinction keeps the reference comparison on a common scale even when the location estimator changes.

Z Additional Two-Dimensional Evaluation Protocols

PDEArena. We use the velocity-field data from the PDEArena benchmark (Gupta & Brandstetter, 2022). The current estimator is evaluated on ten previously used test files, with four trajectories per file. Each trajectory contains fourteen fields: two initialize the recurrence and twelve are predicted from sparse observations. We retain the native 64×64 face-centered velocity representation and the bounded domain. The common closure projects increments onto the kernel of the native discrete divergence, with zero normal boundary increments. It is applied relative to the initial field. All compared estimators use this same closure, bounded location neighborhoods and Euclidean kernel.

The references use 800 training trajectories. Checkpoints at 400 and 1,600 updates are compared on all 100 validation trajectories using complete recurrence; both backbones select 1,600 updates. The

common POD32 basis uses four training trajectories from each of eight training files. History-state estimates use these 32 trajectories. Likelihood parameters are fitted on training trajectories indexed 0 and 16, with index 24 reserved for a training-source diagnostic. These indices are within the training subset and do not refer to test trajectories.

Liquid-metal magnetic fields. The data are the liquid-metal cavity trajectories of Stefanski et al. (2026), generated using the full-induction solver described by Endeve et al. (2025). The reconstruction target consists of the two induced magnetic-field components. We preserve the original configuration split: fourteen configurations for training, two for checkpoint selection, two previous development configurations excluded from this fit and selection, and twelve for evaluation. The evaluation configurations contain all ten magnetic Reynolds number 500 cases and the two field-strength 0.50 cases at Reynolds numbers 100 and 200. Velocities are not supplied to any method.

We sample frames $0, 96, \ldots, 1920$ and use the first two fields to initialize nineteen recurrent steps. A common bilinear resampling from 128×128 to 64×64 preserves the domain endpoints. On the normalized square, the increment projection uses the image of a discrete curl with zero boundary potential. Its normal component is zero on all four walls, while tangential components remain free. This preserves the initial normal field rather than forcing the measured initial boundary residual to zero. The projection specifies the evaluation's discrete constraint and does not reproduce the simulator's finite-element discretization.

Both references select step 1,600 from checkpoints at 400 and 1,600 updates using the two validation configurations. All fourteen training configurations provide the common POD32 basis and history-state estimates. Likelihood fitting uses training indices 0 and 13; index 6 is reserved for a training-source diagnostic.

Shared observations and reporting. Each evaluation uses 64 sensors and five conditions: accurate locations, one-cell location uncertainty, location uncertainty with one-percent observation noise, a missing second component, and location uncertainty with that component missing. AA, AB, BB and the GP moment comparator receive identical observations and location priors, use the same candidate opportunities, and maintain their own recurrent histories. The projected neural reference is reported separately. Training-only checks verify common constraints and agreement between the correction methods when locations are known, including missing-component observations.

The observation seeds are 95100 for PDEArena and 96100 for the magnetic-field evaluation. Means are computed within physical units before aggregation; backbones are reported separately. Partial PDEArena aggregates require all eight tasks of a source file, and partial magnetic aggregates require both backbones of a configuration. The full plans contain 80 and 24 tasks, respectively. These evaluate the frozen current method on data sources previously used in the project. Grid representations, rollout lengths and targets differ across panels, so absolute errors are compared within each panel.

The projected neural reference alone has mean NRMSE 0.869651 and 0.814135 on PDEArena, and 0.464045 and 0.414684 on magnetic fields (FNO and LocalNO, respectively). Table 37 reports the four correction methods.