SafeAdapt: Selective Deployment of Local Corrections to Reduce Harmful Updates in Spatial Soil Carbon Prediction

Abstract

Local corrections can improve average prediction error while degrading some target regions. We study when a few-shot residual correction should replace a shared predictor for modeled lateral soil organic carbon. SafeAdapt combines a frozen source model, source-validation selection of correction amplitude, and a support-only empirical lower-confidence-bound (LCB) gate with source fallback. A harmful update increases held-out regional RMSE relative to the source. Across 33,620 Daily Erosion Project (DEP) episodes, selected-amplitude gating lowers harmful-update incidence from 41.56% to 8.04% at 24.35% update coverage and reduces mean RMSE by 0.89%. Method-aligned European Soil Data Centre (ESDAC) tests show similar harm-coverage trade-offs, although their mean-error gains remain uncertain. A separate ten-seed extension adds selected-candidate controls at matched update counts: LCB lowers conditional harm versus random expectation by 3.90-7.57 percentage points across DEP partitions and 5.34-9.91 points across Global2019 partitions, with exploratory paired site-bootstrap intervals below zero. Advantages over point-gain ranking are smaller and setting-dependent; European equal-coverage harm differences remain uncertain, and fixed shrinkage remains competitive. These results support selective deployment as a practical accuracy-harm-coverage trade-off, rather than a universally optimal correction or a statistical safety guarantee.

1 Introduction

A locally fitted correction is not automatically a better prediction. A shared spatial model may capture broad environmental relationships while leaving regional residual error, but a few local labels may support a correction that fails elsewhere in the same region. The practical decision is therefore not only how to adapt, but whether to replace the existing predictor. A small average gain can be unattractive when many regions become less accurate.

Local adaptation constructs a candidate, while negative-transfer research motivates avoiding inappropriate changes (Wang et al., 2019). Selective prediction considers prediction and rejection jointly (Geifman and El-Yaniv, 2019); here rejection still returns a prediction, using the unchanged source. Sparse, spatially dependent support makes this decision difficult (Roberts et al., 2017). Evaluation must distinguish candidate quality, correction amplitude, and acceptance: a stronger candidate, smaller adjustment, or lower update frequency can otherwise be mistaken for a more informative gate.

SafeAdapt separates these decisions. Residual Ridge fitting uses frozen source representations; source-validation episodes select the correction amplitude. A support-only empirical lower-confidence-bound (LCB) gate assesses improvement on the supplied target scale and retains the source when local evidence is insufficient. No target-query label selects the amplitude or acceptance decision. Correction-versus-fallback is explicit, allowing shrinkage and selective deployment to be evaluated separately.

We report mean RMSE gain, update coverage C, all-episode harmful-update incidence H, and conditional harm $H_{\text{upd}} = H/C$. The identity $H = CH_{\text{upd}}$ distinguishes avoiding updates from selecting useful ones. Equal-coverage controls hold the candidate and update count fixed. This is an evaluation decomposition, not a safety certificate.

Our contributions are the correction-versus-fallback formulation, a reproducible shrinkage-and-gating procedure, and matched accuracy–harm evaluations. DEP provides the primary study; a method-aligned ESDAC extension examines two additional modeled carbon products using separately fitted source models. The extension shares the central adaptation mechanism, not every data-preparation or validation detail. Earlier original-feature Ridge/leave-one-out studies remain separate supporting evidence. All studies use model-derived targets rather than field measurements.

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.

Accordingly, we ask whether selective deployment improves on the source, how error and harm trade off, and whether acceptance adds information beyond reduced coverage. Sections 5.1, 5.2 and 5.4 report primary DEP results; Section 5.5 reports the five-seed ESDAC extension. Sections 5.3 and 5.6 add a separate ten-seed study (Appendix F). Earlier different-protocol studies remain in Appendix D.

2 Application and decision problem

2.1 Spatial soil carbon prediction

DEP supplies the hillslope-level setting. Its target is the supplied erosion-related lateral SOC quantity, not soil carbon stock or net ecosystem carbon balance. The archived specification defines a long-term-average annual flux in g C m^{-2} yr^{-1}, with negative values indicating loss. We evaluate the supplied target values rather than reconstructing annual fluxes. The calendar years underlying this aggregate are not identified in the retained result package. Local labels are model-derived, not independent field measurements; lower prediction error therefore measures agreement with the modeled target, not field accuracy or a conservation benefit.

A HUC12 watershed is the DEP adaptation task. Larger HUC6 domains, one-degree spatial grids, or environmental clusters determine source and held-out domains. This is within-product transfer: source parameters are reused in a target watershed while a small residual correction is learned from local labels. The source network remains frozen. Spatial structure enters the prediction task and evaluation partitions, not a spatially explicit gate or physical-process constraint.

The ESDAC extension evaluates European net lateral carbon change (Fendrich et al., 2024) and global gross erosion-induced SOC displacement at the 2019 baseline (Yoon et al., 2025). Each product has its own source predictor and target scale; no DEP-trained predictor is transferred to ESDAC. Five-degree grid regions replace watersheds, with an offset-grid and an environmental holdout. These are different modeled quantities, not additional field observations. Absolute RMSE is never pooled across products; sampling, units, and protocol departures are specified in Appendix E.

2.2 Correction, fallback, and three distinct outcomes

An episode r has labeled support S_r of size k, a disjoint query set Q_r, a source predictor f_0, and a fixed candidate $f_{\lambda,r}$. A support-only decision $A_r \in \{0, 1\}$ selects

$$f_{A,r}(x) = (1 - A_r) f_0(x) + A_r f_{\lambda,r}(x). \tag{1}$$

Prediction coverage is always complete. *Update coverage* is the fraction of episodes with $A_r = 1$. Let $e_r(f)$ be query RMSE on the supplied target scale and $d_r = e_r(f_0) - e_r(f_{\lambda,r})$. Positive d_r means the candidate helps; deployed utility is $A_r d_r$. With the same episode weights,

$$C = \mathbb{E}[A_r], \quad H = \mathbb{E}[A_r \mathbf{1}\{d_r < -\epsilon\}], \quad H_{\text{upd}} = H/C \quad (C > 0), \tag{2}$$

where $\epsilon = 10^{-12}$ is a numerical tolerance, not a practical damage threshold. H counts harmful updates among all episodes; H_{upd} counts harm among accepted updates and is undefined at zero coverage. Harm frequency does not weight degradation severity.

2.3 Relation to adaptation and selective prediction

Site-specific fitting learns a local model; site calibration fine-tunes the source; residual calibration freezes it and fits a correction (Zeng et al., 2026). Table 1 contrasts these choices with SafeAdapt's deployment decision. Transfer learning reuses information across domains or tasks (Pan and Yang, 2010). MAML learns an initialization for adaptation (Finn et al., 2017); ANIL largely reuses a representation (Raghu et al., 2020). SafeAdapt fits residuals on an ordinarily trained, frozen source without meta-learning. These connections motivate feature reuse, not superiority over untested meta-learned candidates.

Selective classification trades prediction coverage against accepted-prediction risk (Geifman and El-Yaniv, 2017; 2019). Here rejection returns the source, and the relevant loss is the update's effect. WILDS motivates held-out-environment evaluation (Koh et al., 2021). Our within-product partitions and product-specific replications do not demonstrate transfer to field observations or a universal cross-product predictor.

Table 1: Spatial adaptation and deployment strategies.
Application context, not a canonical taxonomy. GNN examples from Zeng et al. (2026) are not evaluated baselines.

Taxonomy	Example / reference	How adapted?	When deployed?
One Size Fits All	Process / Data Driven	No local adaptation	Always source
Spatially Explicit	Global-GNN (Zeng et al., 2026)	Shared spatial model	Always source
Site-Specific	SDSA-GNN† (Sharma et al., 2025)	Fit local parameters	Always adapted
Site Calibrated	FTBSC-GNN† (Zeng et al., 2026	Fine-tune source	Always adapted
Residual Site Calibrated	RSC-GNN (Zeng et al., 2026)	Frozen source + residual	Always adapted
Selective Residual Calibration	**SafeAdapt (this work)**	**Residual correction**	**Support-gated adapt or fallback**

† Cites the original KGML framework; the corresponding GNN variant is defined in Zeng et al. (2026).

3 SafeAdapt: Fitting, Shrinkage, and Acceptance

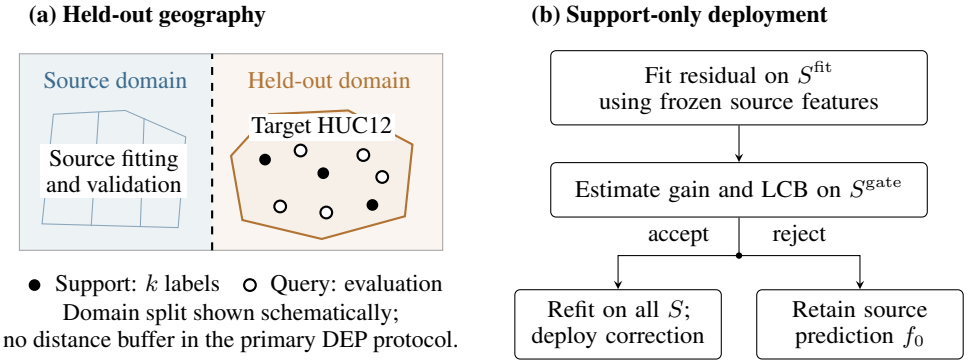


Figure 1: Spatial setting and deployment decision (schematic, not a geographic map). Source data determine the frozen predictor, representation, and validation-selected amplitude. A few labeled records in a held-out watershed provide fitting and gating evidence; disjoint query labels are used only for evaluation. Acceptance deploys the same all-support refit used by the forced control. HUC, grid, and environmental partitions are alternative evaluations, not a spatially explicit gate.

An episode with five labels. At $k = 5$, three support labels fit the evidence candidate and two assess its native-scale gain. Acceptance refits the selected-amplitude procedure on all five; rejection returns the source. The forced policy always uses that same five-label refit. At $k = 1$, the fitting/evidence split is impossible, so the gate retains the source. This illustrates Figure 1, not an additional experiment or a guarantee from two evidence labels.

3.1 Frozen Representation and Residual Candidate

The source network has two width-128 GELU layers and a scalar head. Source-training data determine covariate imputation, standardization, target transformation, and network parameters. For the signed SOC target, define

$$T(y) = \frac{\operatorname{asinh}(y/s) - \mu}{\sigma}, \qquad T^{-1}(z) = s \sinh(\sigma z + \mu), \tag{3}$$

where s is a source-training scale and μ, σ standardize the transformed training targets. Let $h_0(x)$ be the source prediction in this standardized space and $\phi_0(x) \in \mathbb{R}^{128}$ its frozen hidden representation. Writing $z_i = [\phi_0(x_i)^\top, 1]^\top$ and $u_i = T(y_i) - h_0(x_i)$, fit

$$\widehat{\theta}_S = \arg\min_{\theta} \left\{ \sum_{i \in S} (u_i - z_i^\top \theta)^2 + \|\theta\|_2^2 \right\}, \qquad f_{\lambda,S}(x) = T^{-1}\Big(h_0(x) + \lambda z(x)^\top \widehat{\theta}_S \Big). \tag{4}$$

The bias coefficient is penalized, as are the 128 representation coefficients. A dual solve operates on a $k \times k$ system. No source parameter is updated. Shrinkage scales the residual in standardized transformed-target space; it is generally *not* interpolation between native-scale source and corrected predictions. This implementation does not clip signed SOC predictions to a source-range envelope.

The frozen representation reuses features learned across source regions; Ridge regularization makes the fit well defined when coefficients outnumber labels. Amplitude controls the transformed residual's contribution. Because inversion is nonlinear, equal transformed residuals can yield different native-scale changes. Acceptance evidence must therefore be evaluated after inversion to align with reported RMSE.

3.2 Amplitude selection without target-query labels

The candidate amplitudes are $\Lambda = \{0.25, 0.50, 1.00\}$. For each partition, training seed, and support budget, choose the amplitude with the largest mean source-validation RMSE gain, breaking ties toward the smaller amplitude. The choice is then fixed for every target episode in that condition and budget. All 60 choices were checked against their saved source-validation scores: 47 select 0.25, 11 select 0.50, and two select 1.00.

This selection optimizes source-validation *mean gain*; it is not itself a harmful-update constraint. Two policies use the same selected amplitude and same full-support candidate predictions: *forced* deployment always applies the correction, whereas *gated* deployment may return the source. Fixed-amplitude candidates are reported as controls, including candidates that look better on the test episodes. They are not retrospectively relabeled as the validation-selected policy.

3.3 Support-only evidence and empirical LCB gate

The gate randomly partitions the available support into a fitting subset S^{fit} and an evidence subset S^{gate} using the episode's deterministic seed. The evidence subset receives $\lceil 0.4k \rceil$ labels, subject to at least two fitting and two evidence labels. One-label episodes cannot meet these requirements and always fall back. No query label enters this decision.

Fit the selected-amplitude correction on S^{fit}, evaluate both predictors on S^{gate} after inverse transformation, and calculate

$$\widehat{g} = \frac{\mathrm{RMSE}_{S^{\mathrm{gate}}}(f_0) - \mathrm{RMSE}_{S^{\mathrm{gate}}}(f_{\lambda, S^{\mathrm{fit}}})}{\max(\mathrm{RMSE}_{S^{\mathrm{gate}}}(f_0), 10^{-12})}. \tag{5}$$

Resample evidence rows with replacement 200 times, retaining paired source and candidate errors. The 10th percentile of the bootstrapped relative gains is the empirical lower bound ℓ. The candidate is compared with the source's zero-gain action using the ordered tuple (lower bound, point gain, lower complexity). The minimum relative-gain threshold is zero. This is a fixed empirical acceptance rule; there is no target-query threshold search.

After acceptance, deployment uses the candidate refitted on *all* support labels, shared with the forced policy. Consequently, the evidence fit and deployed fit differ. The gate tests the fitting procedure on held-out support, not the exact final coefficients. Small evidence sets, refitting, and within-watershed dependence prevent interpreting the nominal 90% bootstrap level as a 90% safety guarantee. This differs from risk-control procedures with explicitly stated statistical guarantees (Angelopoulos et al., 2024).

Table 2: Information available to each stage. Target-query outcomes evaluate policies but do not select amplitudes, candidate coefficients, or acceptance.

Stage	Information used	Output
Source fitting	Source training; disjoint validation role for early stopping	Frozen network and transforms
Amplitude selection	Source-validation episodes	One λ per condition and budget
Local evidence	Target support only	Candidate-versus-source decision
Deployment	All target support; frozen λ and decision	Full-support correction or source
Evaluation	Held-out target-query labels	Error, harm, and coverage

3.4 Why equal-coverage controls matter

For a fixed candidate within a stratum, let $c = \mathbb{E}[A]$. Selector utility decomposes as

$$\mathbb{E}[Ad] = c\mathbb{E}[d] + \mathrm{Cov}(A, d). \tag{6}$$

Uniform random selection of exactly the same number of updates has expected utility $c\mathbb{E}[d]$. Its contrast with the gate isolates the association between acceptance and benefit rather than the mechanical effect of lower coverage. This is an evaluation decomposition, not a generalization guarantee.

For intuition, suppose a correction worsens 40 of 100 equally weighted episodes. Accepting a random quarter yields ten harmful updates in expectation: all-episode harm falls from 40% to 10%, while conditional harm stays at 40%. This requires no informative selection. A matched-count control tests update allocation; utility additionally accounts for degradation severity. Lower conditional harm need not imply lower RMSE.

We compare the LCB gate with support point gain, source support-residual magnitude, and exact random-selection expectation. Controls use the same candidate and match its gate's update count within each product/partition/seed/budget stratum; ties use fixed hashes. These are offline matched-count diagnostics, not independently calibrated deployment thresholds. Appendix C retains the primary unshrunk comparison. The separate ten-seed extension additionally evaluates selected and fixed-amplitude candidates with product- and partition-specific contrasts (Section 5.3; Appendix F).

4 Experimental design

4.1 Partitions, episodes, and source fitting

DEP primary evaluation. Outer test domains use split seed 56, reserving approximately 20% of domains. Approximately 15% of the remaining domains are reserved for source validation, varying with seeds 42–46. DEP validation sites have disjoint early-stopping, method-tuning, and policy-validation roles. Environmental holdout uses 12 clusters derived from unlabeled site-level covariates across the dataset, not source-only inductive clustering. No explicit distance buffer is part of this protocol.

The source network is trained for at most 12 epochs with AdamW, learning rate 0.001, batch size 2,048, and early-stopping patience three. DEP inputs are header-designated input columns, excluding hillslope/watershed identifiers, coordinates, and output columns. Standardization uses source-training statistics. The residual model uses the frozen 128-dimensional representation, not the original input columns.

Each held-out HUC12 contributes a matched episode per seed and budget. Random support uses a fixed candidate pool; the disjoint query set has at most 2,048 rows. The partitions contribute 393, 858, and 430 eligible target sites, giving 8,405 episodes per budget and 33,620 overall. Repeated geography means these are not 33,620 independent regions.

ESDAC method-aligned extension. Europe and Global2019 each use three spatial or environmental partitions, five seeds, and $k = 1/5/10/20$, yielding 30 product/partition/seed conditions. The central MLP, frozen-hidden Ridge, amplitude grid, support-only bootstrap gate, and fallback mechanism follow DEP. ESDAC instead uses product-specific inputs and grid regions; source validation is reused for early stopping and amplitude selection. Thus the extension is not an exact replication of DEP's validation-role separation. Appendix E specifies these departures and the systematic raster sampling.

4.2 Comparisons and estimands

Within each product, every matched episode receives equal weight, including one-label fallback. This is neither query-pooled RMSE nor an equal-partition average. We report mean RMSE $\bar{e}$, utility $\bar{e}(f_0) - \bar{e}(f)$, and

$$G(f; f_0) = 100\left(1 - \frac{\bar{e}(f)}{\bar{e}(f_0)}\right). \tag{7}$$

The comparisons separate source-only prediction; forced full correction; full correction with the original LCB gate; forced fixed shrinkage at 0.25 or 0.50; and forced or gated validation-selected

shrinkage. Budget- and partition-specific tables expose differences hidden by the aggregate. Equal-coverage comparisons exclude one-label episodes because all gates abstain and no support ranking is defined there.

4.3 Uncertainty, matching, and study status

The paired intervals use 5,000 percentile bootstrap replicates, separately resampling whole seeds or whole partition/site groups. The latter retain repeated seeds and budgets together. There are only five seeds and 1,681 partition/site groups, with geographic overlap across partitions. Neither analysis accounts jointly for neighboring-site dependence and cross-partition overlap. These are separate exploratory sensitivity intervals, not spatial-block guarantees or simultaneous confidence bounds. Random-control intervals concern the exact random-selection expectation, not realized random-mask variation.

All 15 primary DEP conditions completed. Source/full-correction prediction hashes, support/query/pool identities and counts match the deterministic reference. The shrinkage extension adds 201,720 metric rows, not independent episodes. Amplitudes precede query scoring, but the study follows earlier test inspection and is exploratory. The separate ten-seed study has a different DEP cohort and is not pooled with these estimates (Appendix F).

5 Results

5.1 Q1: Can the combined procedure improve on the source?

Table 3 shows that validation-selected shrinkage with gating lowers mean episode RMSE from 4.930 to 4.886, a 0.890% reduction (mean utility 0.044). The seed-clustered 95% interval is [0.035, 0.050], and the partition/site interval is [0.032, 0.056]; both exclude zero under their respective resampling schemes.

Table 3: Primary DEP comparison over all four budgets ($n = 33,620$ episodes). Utility is source RMSE minus policy RMSE. Harm, update coverage, and conditional harm are percentages; conditional harm is undefined at zero updates. All candidates, including better fixed-amplitude mean-error controls, are retained. Differences and gains use unrounded estimates.

Policy	RMSE	Utility	Harm	Update	Cond. harm
Source only	4.930	+0.000	0.00	0.00	–
Full correction, forced	5.477	-0.547	52.30	100.00	52.30
Full correction, gated	4.997	-0.067	7.45	16.26	45.84
Fixed $\lambda = 0.25$, forced	4.845	+0.085	40.46	100.00	40.46
Fixed $\lambda = 0.50$, forced	4.891	+0.038	44.24	100.00	44.24
Selected λ, forced	4.874	+0.056	41.56	100.00	41.56
Selected λ, gated	4.886	+0.044	8.04	24.35	33.03

This improvement cannot be assigned to the original gate alone. Without shrinkage, forced correction increases mean RMSE to 5.477, and gating still gives 4.997, worse than the source. Candidate amplitude changes the error distribution on which the gate operates. The combined positive result therefore supports the implemented shrinkage-and-selection procedure, not a claim that uncertainty-aware rejection alone solves negative adaptation.

5.2 Q2: What is the accuracy–harm trade-off?

Always applying the same validation-selected correction gives lower mean RMSE (4.874) than gated deployment (4.886), but gating reduces harmful updates from 41.560% to 8.040% at 24.350% update coverage; 33.030% of accepted updates are harmful. Thus gating trades 0.012 native RMSE units for substantially fewer degrading updates, not reliable correction in every accepted region.

The forced fixed $\lambda = 0.250$ control has the lowest aggregate RMSE, 4.845, among these policies, with 40.460% harmful updates. Source-validation amplitude selection does not consistently identify

the test-mean optimum. Nor does the new policy dominate the original unshrunk gate: it improves mean error and accepts more updates, but its all-episode harm is slightly higher, 8.04% versus 7.45%. Any preference among these policies must therefore specify the relative costs of mean error, harmful updates, and missed beneficial corrections.

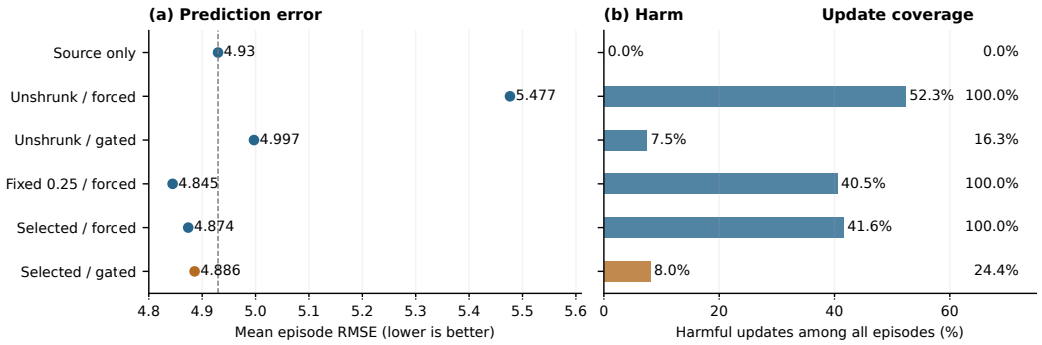


Figure 2: Accuracy, harmful-update incidence, and coverage for six primary policies across all four budgets. These are descriptive point estimates from Table 3, not a threshold sweep or an uncertainty plot. The dashed line is source RMSE. Selected-amplitude gating (orange) occupies a different accuracy–harm trade-off from forced shrinkage; source-only prediction has zero updates and zero harmful updates by construction.

Figure 2 shows why no single column can rank these policies for every deployment objective. Source-only prediction minimizes harmful-update incidence by making no changes. Forced fixed shrinkage has the best aggregate error among the listed controls but changes every episode. Selected-amplitude gating retains a positive mean gain while accepting fewer updates. Its all-budget coverage includes 8,405 one-label episodes that always fall back. Restricting the same policy to $k = 5, 10, 20$ raises coverage to 32.47% and all-episode harm to 10.72%, while conditional harm remains 33.03%. This denominator change is mechanical, not a different fitted policy or additional evidence of gate quality.

5.3 Q3: Does acceptance add value at equal coverage?

The primary unshrunk diagnostics remain in Appendix C. Table 4 instead evaluates the *selected* candidate in a separate ten-seed extension, matching update counts within product/partition/seed/budget at 5/10/20 labels. Paired 95% intervals resample sites within each partition 5,000 times, retaining repeated seeds and budgets; they are exploratory and unadjusted.

Table 4: Ten-seed extension, selected candidate. LCB minus matched comparator in conditional harm (percentage points), estimate [95% interval]. Negative values favor LCB. *n* counts site groups within a partition, not independent cross-partition regions.

Product	Partition	n	Versus random	Versus point gain
DEP	Hydrological	356	-3.90 [-5.72, -1.96]	-2.35 [-3.95, -0.71]
	Spatial grid	796	-5.19 [-6.47, -3.95]	-1.21 [-2.14, -0.29]
	Environmental	835	-7.57 [-8.77, -6.36]	-0.29 [-1.10, 0.52]
Europe	Grid 0	10	+1.08 [-13.65, 18.12]	+5.41 [-3.81, 12.19]
	Grid 2.5	12	+1.02 [-9.92, 15.55]	+0.88 [-8.93, 7.40]
	Environmental	14	-3.31 [-9.08, 3.69]	+4.64 [-0.76, 10.00]
Global2019	Grid 0	93	-9.91 [-12.24, -7.57]	-1.21 [-3.00, 0.51]
	Grid 2.5	91	-9.05 [-11.73, -6.19]	-1.42 [-3.22, 0.30]
	Environmental	99	-5.34 [-8.62, -1.78]	-2.36 [-4.76, -0.08]

All six DEP/Global2019 random-control intervals exclude zero, whereas Europe's do not. Point-gain advantages are smaller and setting-dependent. RMSE contrasts can favor residual-magnitude selection despite its higher harm (Appendix F); no uniform selector dominance is established.

5.4 Q4: How do label budget and partition affect deployment?

Table 5 shows exact source fallback at one label. At 5, 10, and 20 labels, conditional harm declines from 35.77% to 30.48% but remains appreciable. Coverage need not increase monotonically: more labels change both fitting and acceptance evidence.

Table 5: Validation-selected amplitude by DEP label budget. Forced and gated use the same correction. Source RMSE is 4.930; each row has 8,405 episodes. Harm and update entries are percentages.

Labels	Forced RMSE	Gated RMSE	Update	Harm	Cond. harm
1	5.013	4.930	0.00	0.00	—
5	4.858	4.878	33.33	11.92	35.77
10	4.834	4.872	31.14	10.21	32.79
20	4.790	4.864	32.94	10.04	30.48

Table 6: DEP partition-specific means over all budgets. Forced and gated use the selected correction; gain compares gated with source. Partitions reuse geography.

Partition	Source	Forced	Gated	Gain (%)	Cond. harm (%)
Hydrological	5.757	5.744	5.732	0.430	40.68
Spatial grid	5.070	5.048	5.040	0.600	36.42
Environmental	3.893	3.732	3.805	2.280	24.06

The gated point estimate improves in each partition, with conditional harm of 24.06–40.68%. This is heterogeneous deployment behavior, not a causal attribution to a physical driver: site composition, source fit, and support representativeness differ.

5.5 Q5: Does the deployment behavior extend to ESDAC?

The method-aligned extension evaluates 720 European and 5,660 Global2019 matched episodes; all 30 product/partition/seed conditions completed. Each product uses separately trained MLP sources, frozen-hidden Ridge, and selected-shrinkage forced/gated comparisons. These are within-product tests, not DEP-predictor transfer or full-resolution global validation.

Table 7: ESDAC selected-shrinkage extension, all four budgets. H_f/H_g: forced/gated all-episode harm; C_g: gated update coverage; $H_{\text{upd},g}$: harm among accepted updates. Percentage denominators and absolute RMSE scales are product-specific.

Product	Source RMSE	Gated RMSE	Gain (%)	H_f (%)	H_g (%)	C_g (%)	$H_{\text{upd},g}$ (%)
Europe	0.817	0.816	0.127	50.14	10.83	22.92	47.27
Global2019	0.453	0.451	0.445	34.68	7.00	29.98	23.34

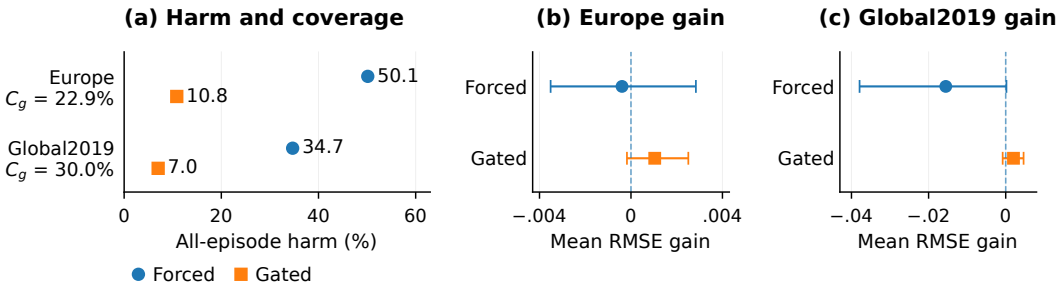


Figure 3: ESDAC selected-shrinkage trade-offs. (a) All-episode harm; C_g is gated coverage (forced coverage is 100%). (b–c) Source-minus-policy mean RMSE and reported 95% intervals on separate product-specific supplied scales. All intervals include zero. Appendix E's exploratory nominal-site grouping is not validated spatial-block uncertainty; these intervals do not test the forced–gated difference.

Table 7 and Figure 3 primarily support fewer harmful updates at reduced coverage. Gated mean RMSE gains are small (0.13% for Europe and 0.45% for Global2019), and both reported intervals include zero, so consistent mean-error improvement is not established. Conditional harm remains 47.27% and 23.34%. Fixed shrinkage can yield lower error, while unshrunk gating remains worse than source. Thus ESDAC supports an accuracy-harm-coverage trade-off rather than universal improvement; Appendix E reports the full policies and uncertainty, and Appendix D keeps the earlier Ridge/LOO study separate.

5.6 TEN-SEED EXTENSION AND OPERATION-LEVEL COST

A separate implementation study with seeds 42-51 completes 90 product/partition/seed conditions and evaluates 79,480 DEP, 1,440 European and 11,320 Global2019 episodes per policy. The DEP cohort differs from the primary study; these results are not pooled. Selected gating's mean seed-specific RMSE reduction is 0.318% [0.226, 0.409] for DEP, 0.018% [-0.136, 0.173] for Europe and 0.342% [-0.265, 0.948] for Global2019. These reported 95% intervals are descriptive normal-approximation seed summaries, not spatial-bootstrap intervals (Appendix F).

Fixed-amplitude gated controls remain competitive: fixed 0.25 has lower pooled RMSE than selected gating on all three products, at different coverages. At 20 labels, bootstrap-gate median-of-medians wall times are 5.16/5.07/5.15 ms for DEP/Europe/Global2019 on an RTX 2000 Ada GPU. These are component timings with cached representations, not end-to-end latency; Appendix F gives the protocol, all controls and timing breakdown.

6 Discussion and limitations

Candidate quality and gate information are distinct. Shrinkage changes the error distribution on which acceptance operates. The primary DEP result supports regularized correction with selective deployment, not an optimal amplitude selector. The ten-seed matched-count controls establish lower conditional harm than random for selected candidates in DEP and Global2019, but improvements over point gain are smaller and setting-dependent. Europe provides no established matched-count harm advantage. Lower conditional harm need not imply lower RMSE.

Risk-averse does not mean guaranteed safe. The gate neither optimizes a tail-risk functional nor enforces a prescribed harmful-update probability. Primary-cohort conditional harm remains 33.03% for DEP, 47.27% for Europe and 23.34% for Global2019. A large error can outweigh many small gains; fallback can also reject beneficial corrections. All-budget coverage includes mechanically abstaining one-label episodes.

Spatial and protocol boundaries. All targets are model-derived; no field accuracy or management benefit is established. The products differ in target semantics, sampled resolution and geography; no explicit distance buffer is imposed. Environmental domains use unlabeled target covariates, and overlapping partitions do not create independent geographies. ESDAC reuses validation for stopping and amplitude choice. The ten-seed implementation also differs from primary DEP in cohort, target-scale and acceptance-boundary details (Appendix F).

Exploratory uncertainty. The studies follow earlier test inspection, not untouched confirmation. The older ESDAC nominal-site grouping has cross-partition ambiguity; new contrasts instead resample within each partition, but still do not account for neighboring-site dependence. Intervals are pointwise and unadjusted. Cross-seed variability and spatial generalization are different questions, and no equal-coverage fixed-versus-selected amplitude advantage is asserted.

7 Conclusion

SafeAdapt treats local calibration as both fitting and selective deployment. In the primary DEP study, selected shrinkage with support-only gating reduces harmful-update incidence from 41.56% to 8.04% while improving mean RMSE by 0.89%. A separate ten-seed extension supports informative update allocation beyond matched random selection for DEP and Global2019, with smaller, setting-dependent gains over point gain. The evidence supports explicit source fallback as a practical accuracy-harm-coverage trade-off, while leaving stronger candidate selection and calibrated risk guarantees open.

Reproducibility and responsible interpretation. Primary artifacts and new summary exports are kept separate. Appendix F documents the extended cohort, controls, uncertainty and timing scope. These modeled-data results make no field-management or carbon-credit recommendation; original providers' data terms apply.

AI Use Statement.

Generative AI tools were used for literature retrieval, language polishing, and assistance with experimental code implementation, debugging, and revision. All AI-assisted content and code were reviewed and verified by the authors.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

A. N. Fendrich, Philippe Ciais, Panos Panagos, Philippe Martin, Marco Carozzi, Bertrand Guenet, and Emanuele Lugato. Including land management in a European carbon model with lateral transfer to the oceans. *Environmental Research*, 245:118014, 2024. doi: 10.1016/j.envres.2023.118014.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*, volume 70 of PMLR, pp. 1126–1135, 2017. `https://proceedings.mlr.press/v70/finn17a.html`.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017. `https://arxiv.org/abs/1705.08500`.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*, volume 97 of PMLR, pp. 2151–2159, 2019.

B. Gelder, T. Sklenar, D. James, D. Herzmann, R. Cruse, K. Gesch, and J. Laflen. The Daily Erosion Project—daily estimates of water runoff, soil detachment, and erosion. *Earth Surface Processes and Landforms*, 43(5):1105–1117, 2018. doi: 10.1002/esp.4286.

Pang Wei Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, volume 139 of PMLR, pp. 5637–5664, 2021. `https://proceedings.mlr.press/v139/koh21a.html`.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020. `https://arxiv.org/abs/1909.09157`.

David R. Roberts et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40:913–929, 2017. doi: 10.1111/ecog.02881.

Arun Sharma, Majid Farhadloo, Mingzhou Yang, Ruolei Zeng, Subhankar Ghosh, and Shashi Shekhar. Spatial distribution-shift aware knowledge-guided machine learning. *arXiv preprint arXiv:2502.14840*, 2025. `https://arxiv.org/abs/2502.14840`.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

J. H. Yoon et al. Impact of soil erosion on soil organic carbon loss and its implications for carbon neutrality. *Advances in Agronomy*, 191:363–414, 2025. doi: 10.1016/bs.agron.2025.01.005.

Ruolei Zeng, Arun Sharma, Shuai An, Mingzhou Yang, Shengya Zhang, Licheng Liu, David Mulla, and Shashi Shekhar. Towards fine-tuning-based site calibration for knowledge-guided machine learning. In 5th Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE), 2026.

Ruolei Zeng, Eduardo Luquin, David Mulla, and Shashi Shekhar. Residual site-calibration of GNN model for spatially heterogeneous soil organic carbon redistribution. In *Proceedings of the 34th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 4 pages, 2026. doi: 10.1145/3841645.3843378.

H. Zheng, C. Miao, C. Huntingford, P. Tarolli, D. Li, P. Panagos, Y. Yue, P. Borrelli, and K. Van Oost. The impacts of erosion on the carbon cycle. *Reviews of Geophysics*, 63:e2023RG000829, 2025. doi: 10.1029/2023RG000829.

Appendices

A IMPLEMENTATION AND INFORMATION BOUNDARIES

Source and candidate. The source optimizer is AdamW with weight decay 10^{-5}. Imputation and scaling are fitted using source-training observations only. The signed-asinh scale s is the median absolute source-training target, with a standard-deviation fallback and then a fallback to one if necessary. The transformed target is standardized using its source-training mean and standard deviation. Signed SOC predictions are inverse-transformed without the physical clipping used for other target types in the broader codebase. This distinction matters: finite values can still be very inaccurate after inversion.

The local residual uses the final 128-dimensional hidden representation and an appended constant. The Ridge penalty is one on every coefficient, including the constant. The equivalent dual expression is $\widehat{\theta} = Z^\top (ZZ^\top + I)^{-1} u$. It is computationally convenient when the support budget is at most twenty. Multiplying this coefficient vector by λ implements transformed-space shrinkage without refitting the source network. Fixed and validation-selected policies receive identical candidates whenever their amplitudes coincide.

Validation and freezing. The held-out test-domain partition uses seed 56. Source-validation allocation varies with the source-training seed; the saved role assignment separates early stopping, method tuning, and policy validation. The extension evaluates each candidate amplitude on permitted source-validation episodes, chooses the largest mean native-RMSE gain for each partition/seed/budget, and serializes the decision before target-query evaluation. Comparing all saved candidate scores with the recorded selections recovers all sixty choices and their tie rule. This check verifies the recorded choice rule; it does not make the reused target geography untouched.

Acceptance and refitting. For $k \geq 5$, the support partition allocates at least two observations to both fitting and evidence. A fixed episode seed controls the split and the 200 paired bootstrap resamples. The empirical tenth percentile, point gain, and action-complexity ordering jointly decide whether the correction is preferred to zero-gain source fallback. The correction is then refitted using all k labels. The same full-support candidate is evaluated for the forced control. At $k = 1$, only the source action is admissible for the gate, while forced controls still evaluate their one-label fit.

Recorded invariants. The final selected-shrinkage archive contains seven policies on the same 33,620 episode keys, or 235,340 matched policy rows. Local reaggregation reproduces 168 summary cells to floating-point precision. The separate experiment audit records agreement of source/full-correction prediction hashes and support/query identities with the deterministic reference run. Agreement of these recorded invariants does not independently validate the underlying process-model labels or every possible implementation path. The supplied numerical ledger distinguishes reaggregated point estimates from archived DEP bootstrap endpoints.

B COMPLETE BUDGET-SPECIFIC PRIMARY RESULTS

Table 8 retains every candidate and support budget rather than reporting only the selected policy or twenty-label setting. The source mean RMSE is 4.93 in every row block. The all-budget aggregate in the main text gives equal weight to these matched episodes. Excluding one-label fallback changes the selected gate's coverage and marginal harm, but not the outcomes of its accepted updates: over 5/10/20 labels, its update coverage is 32.47%, its marginal harm is 10.72%, and its conditional harm remains 33.03%.

C EQUAL-COVERAGE DERIVATION AND UNCERTAINTY

For a stratum of N equally weighted episodes, let A_i be the observed gate action, d_i the candidate's native-RMSE benefit, and $K = \sum_i A_i$. A uniform exact-K random mask R satisfies $\mathbb{E}[R_i] = K/N$,

Table 8: Complete DEP results by label budget. Each policy–budget entry contains 8,405 episodes. "Selected" refers to source-validation selection of λ, not selection of a test-optimal amplitude.

Labels	Policy	RMSE	Harm (%)	Update (%)
1	Full correction, forced	5.45	54.88	100.00
1	Fixed $\lambda = 0.25$, forced	4.91	46.09	100.00
1	Fixed $\lambda = 0.50$, forced	4.98	48.96	100.00
1	Selected λ, forced	5.01	47.77	100.00
1	Full correction, gated	4.93	0.00	0.00
1	Selected λ, gated	4.93	0.00	0.00
5	Full correction, forced	5.74	53.72	100.00
5	Fixed $\lambda = 0.25$, forced	4.85	40.98	100.00
5	Fixed $\lambda = 0.50$, forced	4.94	45.29	100.00
5	Selected λ, forced	4.86	41.17	100.00
5	Full correction, gated	5.07	11.61	23.68
5	Selected λ, gated	4.88	11.92	33.33
10	Full correction, forced	5.50	52.12	100.00
10	Fixed $\lambda = 0.25$, forced	4.82	38.93	100.00
10	Fixed $\lambda = 0.50$, forced	4.87	43.05	100.00
10	Selected λ, forced	4.83	40.54	100.00
10	Full correction, gated	5.03	9.33	20.61
10	Selected λ, gated	4.87	10.21	31.14
20	Full correction, forced	5.21	48.46	100.00
20	Fixed $\lambda = 0.25$, forced	4.79	35.85	100.00
20	Fixed $\lambda = 0.50$, forced	4.78	39.65	100.00
20	Selected λ, forced	4.79	36.78	100.00
20	Full correction, gated	4.96	8.86	20.74
20	Selected λ, gated	4.86	10.04	32.94

even though its entries are dependent. Therefore

$$U_A - \mathbb{E}[U_R] = \frac{1}{N}\sum_{i=1}^{N} A_i d_i - \frac{K}{N}\frac{1}{N}\sum_{i=1}^{N} d_i \tag{8}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(A_i - K/N)(d_i - \bar{d}). \tag{9}$$

The covariance uses divisor N. With multiple strata, the total contrast is the appropriately weighted sum of within-stratum contrasts, not a covariance computed after ignoring their different coverages. Matching weighted coverage is necessary under unequal episode weights. The random control here is an expectation over binary actions; it is not a blended predictor.

These primary unshrunk diagnostics use 25,215 episodes and 5,465 updates (21.67% coverage). LCB/point-gain/residual/random mean RMSEs are 5.02/5.04/5.16/5.04, and conditional harms are 45.84/45.53/53.60/50.17%. Table 9 shows lower conditional harm than random, but no established utility advantage or advantage over point gain. These primary-cohort results concern the unshrunk candidate only. Appendix F separately reports selected- and fixed-amplitude matched-count contrasts from the ten-seed extension; they are not reassigned to this cohort.

D Earlier component studies and protocol separation

D.1 An earlier original-feature Ridge and leave-one-out protocol

The earlier DEP/ESDAC pipeline fits Ridge on standardized original input columns, with an unpenalized intercept, slope penalty ten, and half-amplitude correction. DEP predictions in that pipeline use source-derived numerical bounds. A native-scale leave-one-out support score is followed by a threshold calibrated on separate regions. The primary study instead uses frozen hidden representations, penalty one including the bias, no signed-SOC clipping, a support-holdout bootstrap gate, and

Table 9: Unshrunk LCB minus matched comparator. Negative conditional-harm differences favor LCB; positive utility differences favor LCB. The two bootstrap schemes are separate sensitivity analyses. Intervals are 95% and unadjusted.

Comparator	Outcome	Difference	Seed-clustered interval
Point gain	Conditional harm (pp)	$+0.31$	$[-0.90, 1.44]$
Random	Conditional harm (pp)	-4.33	$[-5.72, -2.86]$
Residual magnitude	Conditional harm (pp)		[
Point gain	Mean utility	+0.02	[−0.03, 0.09]
Random	Mean utility	+0.02	[−0.01, 0.06]
Residual magnitude	Mean utility	+0.14	[0.05, 0.23]
Comparator	Outcome	Difference	Partition/site interval
Point gain	Conditional harm (pp)	$+0.31$	$[-0.89, 1.56]$
Random	Conditional harm (pp)		[
Point gain	Mean utility	+0.02	[−0.02, 0.07]
Random	Mean utility	+0.02	[−0.01, 0.06]

validation-selected amplitude. These are materially different candidate and selection protocols, not a controlled replacement of a single component.

With twenty labels, the earlier DEP protocol reports reductions in mean regional RMSE of 1.56%, 1.68%, and 4.05% under hydrological, spatial-grid, and environmental partitions. Its matched-random reductions are 2.27%, 0.40%, and 0.33%. Harm drops from 39.03/38.44/29.30% under always adapting to 26.82/19.84/18.51% under its gate, at 73.74/55.64/66.37% update coverage. These are separate region-weighted results, not the all-budget episode-weighted primary means in this manuscript.

The earlier protocol also has negative controls. Always adapting has lower mean error than gating on the spatial and environmental partitions. Environmental ten-label source-relative gain is −1.37% with interval [−21.98, 3.26]%. Transformed-unit scoring and a fixed-zero threshold sometimes outperform their native-scale or calibrated alternatives. Historical Head and low-rank updates also show severe finite errors under sparse support. These observations motivate distinguishing candidate construction, numerical transformation, and acceptance; they do not establish a universal repair by native-scale scoring.

D.2 External soil carbon products

The European ESDAC product concerns net lateral carbon change (Fendrich et al., 2024); the global product concerns gross erosion-induced SOC displacement, with a 2019 baseline (Yoon et al., 2025). Net and gross carbon quantities have different interpretations within the wider erosion–carbon cycle (Zheng et al., 2025). Input inventories contain 40,889 European and 225,409 global rows. Each product retains its own source predictor, target scale, and calibration regions. These are modeled products, not independent field observations or transfers of a DEP-trained predictor.

The paired external study uses a 96-tree ExtraTrees source model, three seeds, five or twenty support labels, and at most 128 query observations per episode. A 50 km arm excludes nearby source-training cells. The original paired cohorts have eleven European and ninety-five global test regions. Bootstrap intervals first average seeds within region and then resample regions; repeated geography and small regional cohorts limit their interpretation.

The global buffered result in Table 10 shows selection value beyond its matched random and uncertainty controls, without an established advantage over always adapting. Harm is 22.11% for the gate versus 29.12% for always adapting, at 85.26% update coverage. At five labels, source-relative intervals cross zero. At twenty labels without the buffer, the point gain is 0.92%, without an established matched-random advantage. The buffer changes source fitting and is not a causal test of why selection works.

The original buffered European gate rejects every update and misses the always-adapt gain of 0.89%, with interval [0.06, 2.58]%. Its zero harm is therefore not evidence of useful selection.

Table 10: Earlier ESDAC global twenty-label result with 50 km source separation. Positive reductions favor the gate. These results belong to the original-feature Ridge/LOO protocol, not the primary selected-shrinkage protocol.

Comparator	RMSE reduction (%)	95% interval
Unchanged source	2.20	
Equal-coverage random	0.44	[0.10, 0.73]
Equal-coverage uncertainty	0.94	[0.07, 1.69]
Always adapt	0.13	[−0.16, 0.48]

On an offset partition, update coverage is 75.76% and gain is an uncertain 0.22%, with interval [−0.65, 0.93]%. In another global fixed-source control, halving correction amplitude improves mean RMSE by 2.07%, with interval [0.63, 3.61]%, underscoring the separate contribution of shrinkage.

Units and temporal interpretation. The European catalogue covers 2010–2020 at approximately 10 km; the published lateral-loss map averages 2010–2017 and labels carbon flux in Gg C yr^{-1}. The global published displacement map uses Mg C yr^{-1} at approximately 220 m. These map units do not establish the encoded scale of the sampled raster values retained in the analysis. No physical-unit conversion is applied to those samples. Relative errors are compared within each product; absolute numerical RMSE is not interpreted as a cross-product physical carbon quantity. The European sampled temporal aggregation remains a metadata limitation.

D.3 Limits of frozen reuse and retrieval

An earlier leave-one-product-out policy is a different method from either within-product gate. Its twenty-label gains are uncertain on DEP and Europe, while it accepts no global updates; its one-label DEP gain is negative, −2.48% with interval [−3.13, −0.98]%. Separate hierarchical retrieval and donor controls do not establish consistent value beyond local regularization. These findings are retained as boundaries on reuse, not as evidence for a universal router or as additional contributions. The present paper's conclusion is restricted to the reported local correction and deployment comparisons.

E Method-aligned ESDAC extension

This appendix concerns the new `esdac_dep_protocol_full.py` run, not Appendix D's earlier ExtraTrees/original-feature Ridge/LOO study. The completion log records 30 of 30 product/partition/seed conditions and no condition errors. "Full" denotes completion of the configured evaluation matrix; it does not mean exhaustive evaluation of every native-resolution ESDAC pixel or identity with every DEP implementation detail.

E.1 Products, sampling, and information boundaries

Targets and inputs. The European target is the sum of the supplied aligned lateral-gain and lateral-loss rasters, interpreted as modeled net lateral carbon change. The global target samples the 2019 gross erosion-induced SOC-displacement raster. The implementation uses nearest native-cell values on a systematic 0.125° global grid; it performs no spatial averaging. Both products use 19 WorldClim bioclimatic covariates and elevation. Rows lacking finite target or covariate values are excluded. Feature and target-transform statistics use source-training data. No physical-unit conversion of the encoded raster values is applied.

Partitions and episodes. Local regions are 5° latitude–longitude cells. The three evaluations use an unshifted grid, a grid shifted by 2.5°, and environmental clusters over unshifted-region covariate profiles. Environmental clustering uses unlabeled profiles across the product; it is not an inductive source-only construction. Test domains use seed 56 (approximately 20%); source-validation domains use seeds 42–46 (approximately 15% of remaining domains). There is no explicit distance buffer. Eligible regions have at least 30 retained rows. A fixed 20-row support pool supplies prefixes of size 1, 5, 10, and 20; the remaining query has at most 2,048 rows. Product-specific summaries contain 720 European and 5,660 global matched episodes, or 6,380 total, not 44,660 independent observations. The latter is the seven-policy row count. Regions and pixels overlap across evaluations.

Source models and local decisions. Each condition fits its own two-layer, width-128 GELU MLP. Defaults are AdamW, learning rate 0.001, weight decay 10^{-5}, at most 12 epochs, batch size 2,048, patience three, and training/validation caps of 200,000/50,000 rows. Frozen-hidden Ridge penalizes all 129 coefficients with penalty one. Source validation selects $\lambda \in \{0.25, 0.50, 1.00\}$ by mean RMSE gain, with smaller amplitude breaking exact ties. The selected candidate and forced control share full-support predictions. The gate uses a support fitting/evidence split, 200 paired bootstraps, and the empirical 10th percentile of relative gain; at $k = 1$ it falls back. Target-query labels do not fit the correction, select its amplitude, or determine acceptance.

Table 11: Alignment and departures between the primary DEP procedure and the inspected ESDAC implementation. "Method-aligned" refers to the common central mechanism, not exact protocol identity.

Component	DEP primary	New ESDAC extension
Source and residual	2×128 GELU MLP; frozen-hidden Ridge, penalty one including bias	Same architectural and residual-model choices; separately fitted per product
Amplitude	Source-validation choice from 0.25/0.50/1.00	Same candidate grid and mean-gain selection objective
Validation roles	Disjoint early-stopping, tuning, and policy roles	Same source-validation regions used for early stopping and amplitude selection
Target scale	Median absolute training target with stated fallbacks	Median *nonzero* absolute training target with fallbacks
Acceptance boundary	Ordered lower bound, point gain, complexity	Accept when empirical lower bound is ≥ 0; no lexicographic fallback tie-break
Local geography	HUC12 tasks; HUC6/grid/environmental domains	5° tasks; unshifted/shifted-grid/environmental domains
Inputs and sampling	DEP input columns and supplied records	Climate/elevation covariates; systematic global raster sampling

E.2 Complete product-level results and uncertainty

Table 12: All seven ESDAC policies over all four budgets. Every European row has 720 episodes; every Global2019 row has 5,660. G is source-relative reduction of mean RMSE; H is all-episode harm; C is update coverage. Percentage columns do not imply a physical comparison across products.

Product	Policy	RMSE	G (%)	H (%)	C (%)	H_{upd} (%)
Europe	Source only	0.82	0.00	0.00	0.00	—
	Unshrunk, forced	1.02	−24.49	70.69	100.00	70.69
	Unshrunk, gated	0.82	−0.32	12.92	18.06	71.54
	Fixed 0.25, forced	0.82	0.14	47.78	100.00	47.78
	Fixed 0.50, forced	0.82	−0.66	55.00	100.00	55.00
	Selected, forced	0.82	−0.05	50.14	100.00	50.14
	Selected, gated	0.82	0.13	10.83	22.92	47.27
Global2019	Source only	0.45	0.00	0.00	0.00	—
	Unshrunk, forced	0.53	−16.20	46.34	100.00	46.34
	Unshrunk, gated	0.46	−1.10	8.25	23.94	34.46
	Fixed 0.25, forced	0.45	1.69	29.31	100.00	29.31
	Fixed 0.50, forced	0.44	2.14	34.65	100.00	34.65
	Selected, forced	0.47	−3.43	34.68	100.00	34.68
	Selected, gated	0.45	0.45	7.00	29.98	23.34

Table 13: ESDAC selected-policy mean utilities and reported 95% percentile intervals, in 10^{-3} native RMSE units (not percentages). Utility is source RMSE minus policy RMSE. The 5,000-replicate intervals use the repair script's nominal-site grouping and are exploratory diagnostics, not validated spatial-block intervals.

Product	Policy	Mean utility	Reported 95% interval
Europe	Selected, forced	−0.39	[−3.51, 2.84]
Europe	Selected, gated	1.04	[−0.18, 2.51]
Global2019	Selected, forced	−15.54	[−37.90, 0.19]
Global2019	Selected, gated	2.02	[−0.79, 4.64]

Interpretation. In this five-seed ESDAC cohort, selected gating lowers mean error and harm in point estimates relative to selected forced deployment. The reported intervals do not test that forced-gated contrast; neither source-relative gated mean-gain interval excludes zero. Fixed shrinkage can have lower mean RMSE, particularly on Global2019. Reduced coverage explains part of the all-episode harm reduction, so these aggregates alone do not establish informative selection.

Uncertainty grouping. The earlier repaired bootstrap resamples nominal site identifiers within each product, carrying repeated seeds, budgets and partition rows. Identical grid-index strings can denote different physical cells in shifted and unshifted partitions. This grouping does not establish unique physical-region clusters or account for neighboring dependence and overlap. These intervals remain exploratory, not spatial safety guarantees.

Control-output and audit boundaries. The earlier ESDAC control routine matches counts within product/partition/seed/budget but then pools products with different target scales; those pooled absolute-error outputs are not used here. It also concerns the unshrunk candidate. The *separate ten-seed extension* in Appendix F now supplies product- and partition-specific selected-candidate controls and paired site-bootstrap contrasts. These new contrasts do not retroactively validate the earlier cohort, its nominal-site grouping, or a fixed-versus-selected amplitude comparison at common coverage.

Units and earlier evidence. Encoded-scale and European temporal-aggregation limitations in Appendix D.2 continue to apply. Absolute errors refer to the retained product-specific target values. The earlier 50 km ExtraTrees/LOO arm remains separate and cannot be reassigned to either MLP/hidden-Ridge study.

F Ten-seed extension: matched controls, stability, and cost

F.1 Scope and relation to the primary study

This extension evaluates the central frozen-MLP/hidden-Ridge/shrinkage/gating procedure with seeds 42–51, three partitions per product, and $k = 1, 5, 10, 20$. The completion record reports all 90 product/partition/seed conditions finished with no condition errors. It adds gated fixed-amplitude controls and selected-candidate, product-specific matched-count contrasts. It is a separate implementation study, not ten seeds appended to the original DEP cohort and not an untouched confirmation set.

Table 14: Episode counts are per policy, over all four budgets. The extension's first-five-seed subset is not the primary DEP cohort. No counts are pooled across studies.

Product	Primary (5 seeds)	Extension (first 5)	Extension (10)
DEP	33,620	39,740	79,480
Europe	720	720	1,440
Global2019	5,660	5,660	11,320

Protocol departures. The primary DEP study has 1,681 partition/site groups; this extension has 1,987: 356 hydrological, 796 spatial-grid, and 835 environmental tasks. Europe has 10/12/14 groups and Global2019 has 93/91/99 in unshifted-grid/offset-grid/environmental evaluations. These are partition-specific groups, not mutually independent physical regions. The extension reuses the source-network architecture, hidden-feature Ridge penalty, amplitude grid and support-only evidence mechanism, but does not establish equality of episode hashes with the primary study.

The archived implementation uses the median *nonzero* absolute training target for the asinh scale and accepts when the empirical lower bound is ≥ 0; the primary DEP description includes zero targets in its median and uses an ordered boundary rule. DEP validation is allocated to early-stopping, amplitude-tuning and policy roles, with a fallback to the full validation pool if an eligible role is empty. Whether those fallbacks activated is not resolved by the summary exports. ESDAC reuses source validation for early stopping and amplitude tuning. These differences preclude attributing the change in DEP mean RMSE solely to additional seeds.

Matching and inference. For each candidate, selectors receive identical full-support candidate outcomes. They match the observed gate's exact update count separately within product/partition/seed/budget, excluding $k = 1$. Candidate amplitudes are held fixed *within* a comparison; different candidate families need not have equal coverage. The reported 5,000-replicate paired bootstrap resamples whole sites *within each partition*, carrying repeated seeds and budgets together. It retains observed selector masks rather than reranking in each resample; exact update counts are matched in the original strata, not reimposed after bootstrap resampling. Random-control intervals refer to the exact random-selection expectation, not realized random-mask variation.

Interpretation. All new intervals are exploratory and pointwise, without multiplicity correction. Within-partition grouping avoids combining identically named shifted-grid cells, but does not model neighboring-region dependence. Ten-seed intervals in Table 16 use the exported normal approximation, $\bar{G} \pm 1.96 s_G/\sqrt{10}$, not a t interval, a spatial bootstrap, or ten independent geographic replications. Tables retain unfavorable controls rather than selecting a test-optimal amplitude. Numerical exports and the inspected implementation accompany the revision; the present update does not claim an independent raw-episode rerun.

F.2 Complete product-level outcomes and seed variability

Table 15: Ten-seed extension, all four budgets. G is the reduction of the *pooled mean episode RMSE* (Eq. 7); H, C, and H_{upd} are percentages. Source means are 5.985 (DEP), 0.822 (Europe), and 0.454 (Global2019). Cohort sizes are in Table 14; absolute RMSE is product-specific.

Product	Policy	RMSE	G (%)	H	C	H_{upd}
DEP	Source only	5.985	0.000	0.00	0.00	—
	Unshrunk, forced	6.873	-14.841	55.72	100.00	55.72
	Unshrunk, gated	6.117	-2.206	7.93	15.64	50.72
	Fixed $\lambda = 0.25$, forced	5.937	0.809	39.45	100.00	39.45
	Fixed $\lambda = 0.25$, gated	5.959	0.436	7.73	24.22	31.90
	Fixed $\lambda = 0.50$, forced	6.029	-0.730	44.81	100.00	44.81
	Fixed $\lambda = 0.50$, gated	5.979	0.109	7.89	21.07	37.47
	Selected, forced	5.962	0.384	40.97	100.00	40.97
	Selected, gated	5.966	0.318	7.71	23.09	33.40
Europe	Source only	0.822	0.000	0.00	0.00	—
	Unshrunk, forced	1.194	-45.229	71.11	100.00	71.11
	Unshrunk, gated	0.841	-2.204	13.06	18.61	70.15
	Fixed $\lambda = 0.25$, forced	0.823	-0.044	49.24	100.00	49.24
	Fixed $\lambda = 0.25$, gated	0.822	0.048	12.01	24.10	49.86
	Fixed $\lambda = 0.50$, forced	0.836	-1.703	56.04	100.00	56.04
	Fixed $\lambda = 0.50$, gated	0.823	-0.131	11.39	22.08	51.57
	Selected, forced	0.828	-0.658	52.57	100.00	52.57
	Selected, gated	0.822	0.018	11.32	23.54	48.08
Global2019	Source only	0.454	0.000	0.00	0.00	—
	Unshrunk, forced	0.515	-13.535	45.69	100.00	45.69
	Unshrunk, gated	0.459	-1.158	8.19	24.36	33.61
	Fixed $\lambda = 0.25$, forced	0.447	1.640	29.37	100.00	29.37
	Fixed $\lambda = 0.25$, gated	0.450	0.914	5.93	33.58	17.65
	Fixed $\lambda = 0.50$, forced	0.444	2.241	34.24	100.00	34.24
	Fixed $\lambda = 0.50$, gated	0.448	1.287	6.73	30.64	21.97
	Selected, forced	0.464	-2.174	34.29	100.00	34.29
	Selected, gated	0.452	0.339	6.71	30.27	22.18

Table 16: Seed variability of source-relative RMSE reduction (%). Means average the ten seed-specific G values and can differ slightly from Table 15's pooled G. The intervals are descriptive normal-approximation summaries.

Product	Policy	Mean	SD	Reported 95% interval
DEP	Fixed 0.25, gated	0.436	0.144	[0.346, 0.525]
DEP	Fixed 0.50, gated	0.108	0.264	[−0.055, 0.272]
DEP	Selected, gated	0.318	0.148	[0.226, 0.409]
Europe	Fixed 0.25, gated	0.048	0.065	[0.008, 0.089]
Europe	Fixed 0.50, gated	-0.130	0.399	[−0.377, 0.118]
Europe	Selected, gated	0.018	0.249	[−0.136, 0.173]
Global2019	Fixed 0.25, gated	0.915	0.174	[0.806, 1.023]
Global2019	Fixed 0.50, gated	1.288	0.411	[1.033, 1.543]
Global2019	Selected, gated	0.342	0.978	[−0.265, 0.948]

The selected gate retains a positive DEP mean gain across this seed diagnostic, whereas its European and Global2019 intervals include zero. Gated fixed $\lambda = 0.25$ has lower pooled RMSE than selected gating on all three products, at different coverages; no paired, equal-coverage *cross-amplitude* superiority is inferred. Forced fixed 0.25 (DEP) and forced fixed 0.50 (Global2019) have still lower mean error but update every episode. These results favor reporting an accuracy–harm–coverage trade-off rather than an optimal amplitude selector.

F.3 Selected-amplitude contrasts at matched update counts

Table 17: Selected candidate: LCB minus matched comparator, estimate [95% paired site-bootstrap interval]. Negative values favor LCB in both columns. RMSE differences use 10^{-3} product-specific target units; harm differences are percentage points. No cross-product RMSE pooling.

Comparator	Δ RMSE $\times 10^3$	ΔH_{upd} (pp)
DEP: Hydrological (356 site groups)		
Point gain	−3.606 [−18.244, 11.941]	−2.35 [−3.95, −0.71]
Residual magnitude	+16.068 [−10.495, 42.860]	−1.35 [−4.53, 1.87]
Random expectation	−20.834 [−37.666, −4.934]	−3.90 [−5.72, −1.96]
DEP: Spatial grid (796 site groups)		
Point gain	+0.281 [−6.946, 7.824]	−1.21 [−2.14, −0.29]
Residual magnitude	+17.654 [2.589, 33.800]	−5.02 [−7.27, −2.77]
Random expectation	−8.969 [−18.622, 1.008]	−5.19 [−6.47, −3.95]
DEP: Environmental (835 site groups)		
Point gain	+2.296 [−6.655, 10.438]	−0.29 [−1.10, 0.52]
Residual magnitude	−44.198 [−71.163, −20.906]	−7.71 [−9.62, −5.83]
Random expectation	−18.819 [−29.698, −7.607]	−7.57 [−8.77, −6.36]
Europe: Grid 0 (10 site groups)		
Point gain	+0.919 [−0.042, 2.260]	+5.41 [−3.81, 12.19]
Residual magnitude	−4.281 [−11.997, 0.266]	+4.05 [−16.91, 24.54]
Random expectation	−0.967 [−2.445, 0.283]	+1.08 [−13.65, 18.12]
Europe: Grid 2.5 (12 site groups)		
Point gain	+1.654 [−0.308, 4.215]	+0.88 [−8.93, 7.40]
Residual magnitude	+1.161 [−0.251, 2.898]	+14.04 [−7.20, 33.96]
Random expectation	+1.739 [−0.560, 4.648]	+1.02 [−9.92, 15.55]
Europe: Environmental (14 site groups)		
Point gain	+0.443 [−0.346, 1.341]	+4.64 [−0.76, 10.00]
Residual magnitude	−0.401 [−2.183, 1.160]	+3.97 [−8.87, 17.28]
Random expectation	−1.230 [−2.955, −0.088]	−3.31 [−9.08, 3.69]
Global2019: Grid 0 (93 site groups)		
Point gain	−1.571 [−4.660, 0.453]	−1.21 [−3.00, 0.51]
Residual magnitude	+3.942 [−0.023, 8.733]	−9.67 [−15.70, −3.77]
Random expectation	−2.609 [−4.937, −0.990]	−9.91 [−12.24, −7.57]
Global2019: Grid 2.5 (91 site groups)		
Point gain	-1.065 [-2.397, -4.42×10^{-5}]	−1.42 [−3.22, 0.30]
Residual magnitude	+1.961 [0.253, 3.787]	−12.17 [−18.73, −6.11]
Random expectation	−1.482 [−2.757, −0.531]	−9.05 [−11.73, −6.19]
Global2019: Environmental (99 site groups)		
Point gain	−9.248 [−24.636, 1.177]	−2.36 [−4.76, −0.08]
Residual magnitude	−9.616 [−28.749, 4.745]	−1.57 [−7.25, 4.31]
Random expectation	−5.494 [−19.877, 4.635]	−5.34 [−8.62, −1.78]

What is established by these contrasts? Against matched random expectation, the conditional-harm intervals are below zero in all six DEP and Global2019 partitions, but not in Europe. Against point gain, the conditional-harm intervals exclude zero for DEP hydrological and spatial-grid settings and Global2019 environmental holdout; the other intervals include zero. The Global2019 offset-grid RMSE endpoint is extremely close to zero (-4.42×10^{-8} in native units) and should not be treated as a robust broad advantage. Residual-magnitude selection can have lower RMSE despite greater conditional harm, notably in DEP spatial-grid and Global2019 offset-grid settings. The main conclusion is informative selective allocation in some settings, not uniform selector dominance.

F.4 Fixed-shrinkage controls and the confidence penalty

Table 18: Fixed-amplitude candidates: LCB minus point-gain ranking at the gate's matched count within each candidate. Units and interval construction match Table 17. Each row is a within-candidate comparison, *not* fixed 0.25 versus selected amplitude at common coverage.

Product	Partition	Δ RMSE $\times 10^3$	ΔH_{upd} (pp)
Fixed $\lambda = 0.25$			
DEP	Hydrological	+0.416 [−7.737, 8.479]	−0.54 [−2.13, 1.03]
DEP	Spatial grid	−7.091 [−13.009, −1.465]	−1.30 [−2.23, −0.37]
DEP	Environmental	+0.389 [−2.780, 3.827]	−0.62 [−1.37, 0.12]
Europe	Grid 0	−0.584 [−1.588, 0.032]	+3.37 [−4.13, 11.60]
Europe	Grid 2.5	−0.365 [−1.340, 0.361]	+0.83 [−9.65, 11.06]
Europe	Environmental	+0.165 [−0.276, 0.878]	+0.72 [−6.90, 7.84]
Global2019	Grid 0	−1.595 [−2.954, −0.531]	−1.69 [−3.52, -4.99×10^{-3}]
Global2019	Grid 2.5	−1.290 [−2.671, −0.238]	−3.19 [−5.24, −1.31]
Global2019	Environmental	−0.558 [−1.177, 0.032]	−2.66 [−4.39, −1.08]
Fixed $\lambda = 0.50$			
DEP	Hydrological	−11.041 [−32.649, 8.709]	−2.32 [−4.26, −0.49]
DEP	Spatial grid	−3.345 [−21.458, 19.119]	−1.16 [−2.34, −0.05]
DEP	Environmental	−2.029 [−9.248, 5.334]	−0.60 [−1.47, 0.28]
Europe	Grid 0	+0.259 [−0.094, 0.664]	+2.38 [−4.56, 11.34]
Europe	Grid 2.5	−0.002 [−0.616, 0.491]	−4.85 [−11.95, 1.04]
Europe	Environmental	+0.434 [−0.060, 1.076]	+3.82 [−2.32, 9.27]
Global2019	Grid 0	−1.590 [−4.447, 1.225]	+0.35 [−1.57, 2.31]
Global2019	Grid 2.5	−0.646 [−2.479, 0.781]	−2.90 [−5.08, −0.94]
Global2019	Environmental	−0.667 [−2.643, 1.366]	−3.09 [−5.18, −1.23]

Candidate strength is a separate issue. Table 15 shows that fixed shrinkage remains competitive with validation-selected amplitude. Table 18 asks a different question: whether the confidence penalty improves allocation for a *fixed* candidate, at the same count as point-gain ranking. Some fixed-0.25 settings improve in both metrics, whereas many intervals cross zero. Choosing the amplitude after viewing Table 15 would be a different, test-selected policy, not evidence that source validation selected it.

Other controls. For fixed 0.25, conditional harm is lower than matched random in all DEP and Global2019 partitions under the reported intervals. Relative to residual magnitude, however, fixed-0.25 LCB has worse RMSE in DEP hydrological and all three Global2019 partitions. With selected amplitude, the DEP spatial-grid and Global2019 offset-grid contrasts also show a lower-harm/higher-error trade-off (Table 17). These unfavorable comparisons are retained in the accompanying complete 324-row contrast export, which includes all four candidate families, three comparators, and three outcomes.

Boundary on the new evidence. Matching is exact among selectors for a candidate. The observed gate coverages differ across candidates. Neither Table 15 nor Table 18 establishes a paired fixed-0.25 versus selected-amplitude advantage at common coverage, and no such result is asserted. The evidence narrows the role of amplitude selection to one implementable candidate-construction choice, while supporting evaluation of acceptance separately from correction size.

F.5 Coverage diagnostics and operation-level runtime

Table 19: Selected-candidate coverage sweep, $k = 5, 10, 20$. Each selector takes the same rounded count per product/partition/seed/budget; the actual coverage is reported. Product-level aggregates across partitions are *descriptive*, not independent replications. LCB here denotes top-score ranking, not the zero-bound deployment rule. PG: point gain; R: exact random expectation.

Product	C (%)	RMSE LCB	RMSE PG	H_{upd} LCB	PG	R
DEP	10.1	5.9702	5.9745	30.18	34.02	39.68
DEP	20.0	5.9625	5.9661	31.53	34.08	39.67
DEP	30.0	5.9584	5.9622	33.46	34.69	39.67
DEP	40.0	5.9618	5.9570	35.37	35.08	39.67
DEP	50.0	5.9634	5.9518	37.03	35.90	39.67
Europe	8.3	0.8218	0.8216	43.33	36.67	49.29
Europe	19.4	0.8220	0.8213	48.10	42.86	48.84
Europe	30.6	0.8230	0.8210	47.27	43.03	48.77
Europe	41.7	0.8230	0.8212	48.44	45.11	48.70
Europe	50.0	0.8231	0.8225	49.63	46.11	48.80
Global2019	9.9	0.4529	0.4529	19.17	20.60	31.14
Global2019	20.1	0.4529	0.4536	20.41	23.10	31.09
Global2019	30.0	0.4540	0.4549	21.57	24.63	31.11
Global2019	39.9	0.4541	0.4586	23.33	24.90	31.12
Global2019	50.2	0.4554	0.4596	25.31	25.26	31.10

LCB is not uniformly preferable across coverage. On DEP, its descriptive error and conditional harm are lower than point gain at the 10–30% targets, but not at 40–50%. Europe shows no consistent LCB advantage; Global2019 has lower LCB error at several intermediate targets but loses its pooled source-relative gain at higher coverage. The sweep is diagnostic, not a test-set choice of deployment threshold.

Table 20: Operation-level wall time (ms), median of per-episode medians at $k = 20$. Measurements per operation: DEP 360, Europe 340, Global2019 360. The completion log identifies an NVIDIA RTX 2000 Ada Generation GPU (15.6 GiB reported VRAM).

Operation	DEP	Europe	Global2019
Source query inference	0.229	0.233	0.231
Support feature extraction	0.210	0.208	0.210
Ridge fit (cached features)	0.069	0.067	0.068
Bootstrap gate	5.162	5.069	5.150
Full refit + final prediction	0.097	0.164	0.128

The archived timing routine uses a deterministic prefix of eligible regions, one warmup and three measured repetitions by default, with CUDA synchronization around wall-clock timing. Ridge, gating and final correction operate on cached source representations; source query inference and support feature extraction are measured separately. The gate measurement includes its evidence fit and 200 paired bootstrap resamples. These are component timings, not end-to-end latency, not a sum of medians, and not a measurement of source training, raster preparation, or amplitude search. The reported gate cost is about 5.1 ms in this setting; it does not establish low overhead relative to source inference for every deployment workload.