When Is Source Information Worth Retaining? Evidence-Gated Selective Transfer for Sparse Geothermal Profiles

Anonymous authors
Paper under double-blind review

Abstract

Sparse geothermal temperature logs often leave long depth intervals unobserved, making profile reconstruction difficult when only a few target measurements are available. Target-only methods avoid cross-well mismatch, but sparse anchors often constrain the large-scale trend more strongly than the nonlinear structure inside a wide gap. Source-assisted methods can recover some of that structure from other wells, but only when source and target profiles are sufficiently compatible. With few target anchors, transfer can be selected too optimistically if the same observations are used both to calibrate a source correction and to judge its reliability. We introduce Evidence-Gated Selective Transfer (ESGT), which separates these two roles. Geometry-matched inner pseudo-gaps estimate the correction magnitude, outer held-out anchor blocks evaluate the calibrated correction without reusing their labels, and a detrended source-target shape statistic provides an additional compatibility check. The source correction is retained only when these signals support transfer; otherwise, ESGT uses a target-only reconstruction. We evaluate ESGT under controlled structural shifts and on measured geothermal profiles, with additional report-derived profiles prepared through lightweight multimodal indexing. Across the controlled scenarios, ESGT reduces RMSE from 0.876 for the strongest selective baseline to 0.822 while substantially reducing harmful transfer under source-target mismatch. On the Montana field data, it achieves an RMSE of 0.370°C. Separating calibration from retention is most useful when source relevance varies across targets.

1 Introduction

Sparse subsurface temperature logs are common in geothermal profile analysis, where measurements may be available at only a small number of depths and an internal interval remains unobserved. Many downstream analyses require a continuous temperature-depth profile, so the missing interval must be reconstructed from the available measurements. We study this within-profile reconstruction problem rather than regional temperature mapping or full subsurface modeling: the target well already contains sparse temperature observations, and the goal is to recover the missing temperatures between them. The problem becomes harder as the gap widens or the number of observed anchors decreases, because sparse measurements constrain the large-scale trend more reliably than the nonlinear structure inside the gap. Figure 1 illustrates the setting and the role of auxiliary source information.

For this setting, it is useful to distinguish methods that use only observations from the target well from those that also use auxiliary wells. We call the former target-only methods. Shape-preserving interpolation and Gaussian-process (GP) regression are representative examples because they reconstruct the missing interval directly from the observed target profile (Fritsch & Butland, 1984; Rasmussen & Williams, 2006). They avoid cross-well mismatch, but a small anchor set may contain too little information to recover unresolved structure inside a wide gap. Source-assisted approaches draw on related wells or source domains, following the broader ideas of transfer learning and multi-source regression (Pan & Yang, 2010; Mansour et al., 2008; Wei et al., 2017; Hickey et al., 2024). In geothermal profiles, these auxiliary data can provide temperature-depth structure that is only weakly constrained by the target anchors.

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.

The usefulness of source information, however, varies across targets and gaps. A compatible source can supply structure that the target anchors do not resolve, whereas a mismatched source can increase reconstruction error. A source-assisted predictor must therefore determine both how much of the source-induced correction to apply and whether that correction should be retained for the current target. Deciding whether to retain the correction is difficult when only a small number of target observations are available. In particular, reusing the same target observations both to calibrate a source correction and to assess its reliability can produce overly optimistic evidence for transfer. A correction may agree with the available observations while still degrading reconstruction inside the actual unobserved interval. The central question is therefore whether the available target evidence is sufficient to support retaining source information for the current gap.

ESGT addresses this question by separating correction calibration from source retention. Geometry-matched inner pseudo-gaps estimate the magnitude of the source-induced correction, while outer held-out anchor blocks test the calibrated correction without using their labels in the corresponding calibration step. A detrended source-target shape statistic supplies an additional compatibility check. ESGT keeps the source correction only when these signals support transfer; otherwise, it uses the target-only reconstruction. This design makes source retention as an explicit, target-conditioned decision. The evidence used to calibrate the correction is separated from the evidence used to determine whether it should be applied, while the target-only reconstruction remains available when transfer is not supported. We therefore evaluate both reconstruction error and the quality of the transfer decision, including whether harmful corrections are rejected and beneficial ones are retained. Experiments cover controlled structural shifts, measured geothermal profiles, and additional report-derived profiles organized through lightweight multimodal indexing. Across these settings, selective retention is most useful when the source contains structure that the sparse target anchors cannot resolve on their own.

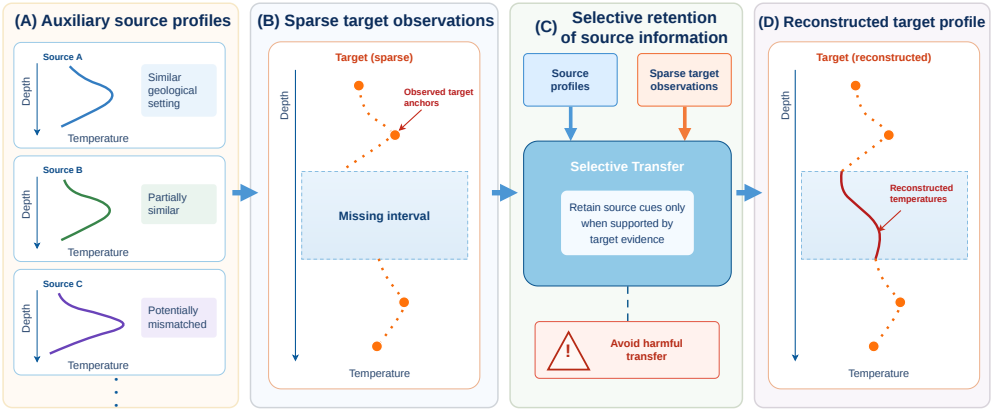


Figure 1: Task overview. Source profiles provide auxiliary temperature-depth structure, the target profile contains sparse observations around an internal missing interval, and the goal is to reconstruct that interval using source information and target anchors.

2 Related Work

Pan and Yang (Pan & Yang, 2010) surveyed transfer learning methods for target tasks with limited target data. Ben-David et al. (Ben-David et al., 2010) related target error to source error and source–target discrepancy, while multi-source studies showed that naive source averaging can fail when sources differ in relevance (Mansour et al., 2008; Hoffman et al., 2017). Kouw and Loog (Kouw & Loog, 2021) reviewed adaptation through sample weighting, representation changes, and inference-time modification. More recent work studies source usefulness through adaptive source use (Hu & Lin, 2025; Deng & Kpotufe, 2025), transferability estimation (Tan et al., 2025), and explicit control of negative transfer (Xu et al., 2026; Sun et al., 2026). These studies consider different transfer settings, but share the question of when source information remains useful under limited target supervision.

For continuous regression, several lines of work are relevant to how source information is represented and adapted. STPS-MTR transfers locally similar source–target segments (Yang et al., 2021), while robust multiple-source regression constructs predictors from mixtures of source distributions (Hoffman et al., 2017). Recent work also studies how source information should be represented or allocated across multiple sources (Ng et al., 2025; Zhang et al., 2025) and how source–target overlap affects transfer in high-dimensional regression (Tahir et al., 2025). Gaussian-process methods model source–target relatedness directly. TC$_{MS}$Stack uses transfer covariance and similarity-linked stacking (Wei et al., 2017), while transfer-kernel learning encodes pairwise domain relatedness in the covariance structure (Wei et al., 2023). RECaST instead recalibrates a fixed source predictor through a random effect (Hickey et al., 2024). Several other methods require additional segmentation, representation-learning, or training components, so we use TC$_{MS}$Stack and RECaST as literature baselines that can be instantiated directly from the source profiles and sparse target anchors available in each episode.

Nested validation is also related to model selection under distribution shift. Domain-generalization and distribution-shift benchmarks emphasized validation across changing environments (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021), while spatial validation and leakage studies examined dependence among reused or correlated observations (Roberts et al., 2017; Kapoor & Narayanan, 2023). Recent domain-adaptation theory has further studied target-risk guarantees and model selection under heterogeneous shift (Wu et al., 2025; Zhou et al., 2026). Learn then Test and conformal risk-control methods calibrated decisions against explicit risk criteria (Angelopoulos et al., 2025; 2024). ESGT uses nested pseudo-gap evidence for a narrower question: whether a source correction should be retained for the current target gap.

In geothermal and well-log applications, classical interpolation and regression-kriging (Fritsch & Butland, 1984; Hengl et al., 2007) have been complemented by physics-informed transfer (Shima et al., 2024), missing-log imputation and reconstruction (Gama et al., 2025; Zhou et al., 2025), continuous thermal-profile construction from incomplete borehole measurements (Mou et al., 2026), regional temperature mapping (Burns et al., 2024; Gordon et al., 2025), and sparse-observation subsurface prediction (O'Malley et al., 2026). These studies establish the broader relevance of reconstruction from incomplete subsurface observations. Our setting focuses on whether information from auxiliary wells should be retained for a particular missing interval when only a small number of target measurements are available.

3 Problem and Evaluation Protocol

For a target profile, let $A = \{(a_j, y_j)\}_{j=1}^{k}$ denote the observed target anchors, where a_j is depth, y_j is the corresponding temperature, and $k = |A|$. Let Q be a disjoint set of query depths with hidden temperatures y_Q, and write $y(x)$ for the target temperature at depth x. The source function f_s is built from the admissible source pool for the corresponding evaluation setting. All methods are evaluated on the same A, Q, and source pool, and y_Q is used only for final scoring and retrospective diagnostics. For a finite depth set S, define mean squared error (MSE) as $\mathrm{MSE}_S(f) = |S|^{-1} \sum_{x \in S} (y(x) - f(x))^2$, the average inner product as $\langle u, v \rangle_S = |S|^{-1} \sum_{x \in S} u(x)v(x)$, and $\|u\|_S^2 = \langle u, u \rangle_S$. The operator $\mathrm{clip}_{[0,1]}$ truncates its argument to $[0, 1]$.

3.1 Measured and simulated data

The primary measured dataset contains 1,518 retained temperature-depth observations from 30 Montana wells in 17 counties, obtained from the Southern Methodist University (SMU) node of the National Geothermal Data System (NGDS) archive (Blackwell et al., 2014). We retain records with documented metre/Celsius units, at least 30 unique depths, and at least 100 m of depth span. Repeated observations are reduced deterministically, and each well contributes at most 61 measured depths. Wells from the target county are excluded from its source pool.

An additional measured set contains geothermal profiles extracted from confidential engineering and geological reports and standardized as depth-referenced numerical records. We report aggregate results for this set.

We select up to four source wells by the similarity of their observed physical depth spans to the target span. For modeling, each profile is represented on the common rescaled depth coordinate $[0, 3000]$

and interpolated with Piecewise Cubic Hermite Interpolating Polynomial (PCHIP). Let f_i denote source profile i and f_s their equal-weight mean. ESGT and the residual/calibration baselines use f_s, while TC$_{MS}$Stack retains the individual f_i. For Montana, we use relative gap widths $0.2, 0.4, 0.6$, relative centers $0.35, 0.5, 0.65$, and anchor budgets $k \in \{5, 10\}$. Each feasible anchor set contains k observations, including the profile endpoints and the nearest observed point on each side of the gap. Query points are existing measurements inside the gap, giving 540 feasible episodes.

The controlled simulator solves one-dimensional steady heat conduction on a one-metre grid and samples 61 observation depths. Shared-structure targets reuse the source layer boundaries and relative conductivities, whereas independent-structure targets redraw the stratigraphy. Near-linear targets use homogeneous conductivity with no volumetric heat generation. Gaussian observation noise has standard deviation 0.5°C. We use 20 scenarios per regime for validation and 50 separate scenarios per regime for final evaluation. Each final scenario contains six gap and anchor-budget episodes, giving 900 simulation episodes in total.

3.2 Aggregation

RMSE is computed at the episode level. For simulation, we average episodes within each scenario, scenarios within each regime, and then give the three regimes equal weight. For Montana, gap settings are averaged within wells, wells within counties, and counties equally. For the report-derived profiles, RMSE is averaged within each profile and then equally across profiles.

4 Evidence-Gated Selective Transfer

ESGT starts from a target-only fallback and adds a source-induced correction whose magnitude is estimated from geometry-matched pseudo-gaps. Held-out target anchors are then used to decide whether that correction should be retained. Appendix E gives the retrospective source-value decomposition used in the analysis.

4.1 Prediction form

For anchor set A, let b_A denote the selected target-only fallback and t_A the source-assisted candidate. Let h_Q be the span of the actual query gap, $\hat{g}_A(h_Q) \in [0, 1]$ the calibrated source-correction coefficient for the full anchor set, and $r_A \in \{0, 1\}$ the source-retention decision. ESGT predicts

$$\hat{y}_A(x) = b_A(x) + r_A \hat{g}_A(h_Q)\,[t_A(x) - b_A(x)]\,. \tag{1}$$

Here b_A is the target-only prediction, $t_A - b_A$ is the source-induced correction, $\hat{g}_A$ sets its magnitude, and r_A determines whether it is applied.

4.2 Target fallback and source candidate

ESGT uses a target-only fallback when source information is not retained. The fallback is chosen from $\mathcal{M} = \{\text{PCHIP}, \text{Target GP}, \text{Trend GP}\}$, which represent three common target-only modeling choices for sparse profiles. PCHIP provides local shape-preserving interpolation, the Target GP models smooth variation directly from the observed target anchors, and the Trend GP combines an affine temperature–depth trend with a residual GP. Both GP models use a Matérn-$3/2$ covariance with length scale 500 on the rescaled depth axis and nominal observation noise 0.5°C. Appendix B gives their explicit forms. The fallback is selected from these candidates using geometry-matched pseudo-gap error on the observed target anchors.

For candidate $m \in \mathcal{M}$, let $m_{A\setminus J}$ be the model refitted after removing internal pseudo-gap block J, and let $\mathcal{J}_A(h_Q)$ contain the feasible blocks for query-gap span h_Q. Blocks receive larger weights when their spans more closely match the query gap. The weighted pseudo-gap loss and selected fallback are

$$\begin{aligned} \mathcal{L}_A(m; h_Q) &= \sum_{J \in \mathcal{J}_A(h_Q)} w_J \,\mathrm{MSE}_J\left(m_{A\setminus J}\right), \\ b_A &= \arg\min_{m \in \mathcal{M}} \mathcal{L}_A(m; h_Q). \end{aligned} \tag{2}$$

Ties are resolved in favor of PCHIP. We refer to b_A as the adaptive target fallback.

Let $\mathcal{G}_A$ denote the Matérn-$3/2$ residual GP fitted on A. The source-assisted candidate at depth x is

$$t_A(x) = f_s(x) + \mathcal{G}_A\{y_A - f_s(A)\}(x), \tag{3}$$

where $y_A = (y_1, \ldots, y_k)^\top$ and $f_s(A) = (f_s(a_1), \ldots, f_s(a_k))^\top$.

4.3 Geometry-matched calibration

For a retained anchor set B, ESGT forms contiguous internal blocks J containing one or two anchors while keeping both endpoints. Let h_J be the span between the retained anchors bracketing J, and let h be the span to be predicted. We weight each block by

$$w_J = \exp\left(-\frac{|\log(h_J/h)|}{\tau}\right), \qquad \tau = 0.5. \tag{4}$$

We retain at most the eight highest-weight blocks and normalize their weights to sum to one. Let y_J be the observed temperature vector on J, and let $b_{B\setminus J}$ and $t_{B\setminus J}$ be the fallback and source candidate refitted after removing J. Define $e_J = y_J - b_{B\setminus J}(J)$ and $d_J = t_{B\setminus J}(J) - b_{B\setminus J}(J)$. The calibration statistics are

$$\begin{aligned} \widehat{C}_B &= \sum_J w_J \langle e_J, d_J\rangle_J, & \widehat{D}_B &= \sum_J w_J \langle d_J, d_J\rangle_J, \\ \hat{g}_B(h) &= \mathrm{clip}_{[0,1]} \frac{\widehat{C}_B}{(1+\rho)\widehat{D}_B}, & \rho &= 0.2. \end{aligned} \tag{5}$$

Here ρ controls shrinkage. If no feasible block exists or the denominator is negligible, the coefficient is set to zero. The same rule applies to any retained anchor set B and prediction span h; setting $B = A$ and $h = h_Q$ gives the coefficient $\hat{g}_A(h_Q)$ used in Equation 1.

4.4 Outer evidence and source gate

Outer pseudo-gap blocks H contain one or two contiguous internal anchors and retain the two profile endpoints. A block is feasible if its removal leaves enough anchors to refit the fallback and form at least one inner calibration block. Let h_H be the span between the retained anchors bracketing H. We use the geometry-matching rule in Equation 4 with $h = h_Q$ and normalize the resulting weights w_H across the retained outer blocks.

For each outer block $H \subset A$, set $B = A\setminus H$ and estimate $\hat{g}_B(h_H)$ using only inner blocks contained in B. Let b_B and t_B be the fallback and source candidate refitted on B. The held-out gain is

$$U_H = \mathrm{MSE}_H(b_B) - \mathrm{MSE}_H\left[b_B + \hat{g}_B(h_H)(t_B - b_B)\right]. \tag{6}$$

A positive U_H means that the calibrated source correction improves on the target fallback for the held-out block. With numerical tolerance $\varepsilon = 10^{-10}$, define

$$\begin{aligned} S_A &= \frac{\sum_H w_H U_H}{\max\left(\sum_H w_H \,\mathrm{MSE}_H(b_B), \delta_{\mathrm{floor}}\right)}, \\ R_A &= \sum_H w_H \mathbf{1}\{U_H > \varepsilon\}. \end{aligned} \tag{7}$$

The denominator floor is $\delta_{\mathrm{floor}} = 0.25$. S_A measures the relative gain magnitude, and R_A measures how consistently the gain is positive across outer blocks.

To compare shape independently of the large-scale trend, we fit an affine function of depth separately to $f_s(A)$ and y_A. Let $\widetilde{f}_A$ and $\widetilde{y}_A$ be the resulting residual vectors. Their cosine similarity is

$$s_A = \frac{\widetilde{f}_A^\top \widetilde{y}_A}{\|\widetilde{f}_A\|_2 \|\widetilde{y}_A\|_2}. \tag{8}$$

If either residual norm is numerically negligible, we set $s_A = 0$. ESGT retains the source correction when

$$r_A = \mathbf{1}\{S_A > \theta_S,\ R_A \geq \theta_R,\ s_A \geq \theta_s\}. \tag{9}$$

The selected thresholds are $(\theta_S, \theta_R, \theta_s) = (0.05, 0.5, 0.70)$. If the gate rejects transfer, $r_A = 0$ and Equation 1 returns the selected target-only fallback.

4.5 END-TO-END PROCEDURE

Algorithm 1 summarizes a target episode. Each outer block is removed before fitting the coefficient used to evaluate it.

Algorithm 1 Evidence-Gated Selective Transfer for one target episode

Require: target anchors A, query depths Q, admissible source profiles $\{f_i\}$, thresholds $\theta_S, \theta_R, \theta_s$
Ensure: prediction $\hat{y}_A(Q)$

```
Select the target fallback b_A by Equation 2.
Build f_s from the admissible source pool and fit the source candidate t_A by Equation 3.
for each feasible outer block H ⊂ A defined in Section 4.4 do
    Set B ← A \ H and refit b_B and t_B on B.
    Construct internal blocks J in B and estimate ĝ_B(h_H) by Equation 5.
    Evaluate the held-out gain U_H by Equation 6.
end for
Aggregate outer evidence into S_A and R_A by Equation 7.
Compute detrended source-target cosine similarity s_A and gate decision r_A by Equation 9.
Refit calibration on all anchors to obtain ĝ_A(h_Q).
return ŷ_A(Q) by Equation 1.
```

5 EXPERIMENTS

5.1 METRICS AND COMPARATORS

For episode i, let b_i be the target fallback, t_i the unshrunk source-assisted candidate, and $\hat{y}_i$ the deployed prediction. Let $z_i \in \{0, 1\}$ indicate source use, with $z_i = 1$ when the method applies a nonzero source correction and $z_i = 0$ when it uses the target fallback alone. Define

$$B_i = \mathrm{MSE}(b_i), \qquad C_i = \mathrm{MSE}(t_i), \qquad D_i = \mathrm{MSE}(\hat{y}_i), \tag{10}$$

where all three MSEs are evaluated on the episode's query set. With numerical tolerance $\varepsilon = 10^{-10}$, the decision metrics are

$$\begin{aligned} \mathrm{Active} &= \frac{1}{N}\sum_i z_i, & \mathrm{Harm} &= \frac{1}{N}\sum_i \mathbf{1}\{D_i - B_i > \varepsilon\}, \\ \mathrm{Recall} &= \frac{\sum_i z_i \mathbf{1}\{B_i - C_i > \varepsilon\}}{\sum_i \mathbf{1}\{B_i - C_i > \varepsilon\}}, & \mathrm{Reject} &= \frac{\sum_i (1 - z_i)\mathbf{1}\{C_i - B_i > \varepsilon\}}{\sum_i \mathbf{1}\{C_i - B_i > \varepsilon\}}. \end{aligned} \tag{11}$$

Active is the fraction of episodes with a nonzero source correction. Harm is the fraction in which the deployed prediction is worse than the target fallback. Recall measures how often retrospectively beneficial source candidates are retained, and Reject measures how often harmful candidates are rejected.

We compare PCHIP, Target GP, Trend GP, the adaptive target fallback, Residual GP, gain calibration, shape-gated calibration, ESGT, and a retrospective oracle coefficient. Residual GP deploys t_A without calibration or gating. Gain calibration applies $\hat{g}_A(h_Q)$ without the outer evidence gate, while shape-gated calibration also requires $s_A \geq \theta_s$. ESGT additionally uses the held-out evidence summarized by S_A and R_A. Appendix B gives the explicit baseline prediction rules. The oracle coefficient minimizes query error over $g \in [0, 1]$ using hidden query labels and is used only for retrospective analysis.

For TC$_{MS}$Stack, we use source–target transfer covariances and similarity-linked stacking with a fixed squared-exponential kernel and anchor-only similarity optimization. The RECaST adapter uses the continuous-outcome Gaussian–Cauchy random-effect likelihood, maximum a posteriori (MAP) fitting, and a plug-in predictive median. Appendix D gives implementation details.

These baselines isolate the main stages of the method. The target-only models define the fallback alternatives, Residual GP exposes the unshrunk source correction, Gain calibration estimates its magnitude, Shape-gated calibration adds a source–target shape check, and ESGT adds held-out outer evidence. Appendix E reports a retrospective decomposition of realized source gain into oracle headroom, calibration shortfall, and retention regret.

Table 1: Controlled simulation comparison, mean episode RMSE in °C. Each regime contains 50 independent scenarios and six gap and anchor-budget episodes per scenario. The oracle is diagnostic rather than deployable.

Method	Shared	Independent	Near linear	Macro
PCHIP	1.234	1.226	1.010	1.157
Target GP	2.831	2.526	2.692	2.683
Trend GP	1.534	1.313	0.571	1.139
Adaptive target fallback	1.170	1.160	0.571	0.967
Residual GP	0.951	2.102	0.919	1.324
RECaST, MAP adapter	1.173	4.593	1.132	2.299
TC$_{MS}$Stack, fixed kernel	1.450	2.490	1.524	1.821
Gain calibration	0.925	1.150	0.577	0.884
Shape-gated calibration	0.930	1.130	**0.568**	0.876
ESGT	**0.860**	**1.035**	0.570	**0.822**
Oracle coefficient, diagnostic	0.750	0.930	0.550	0.743

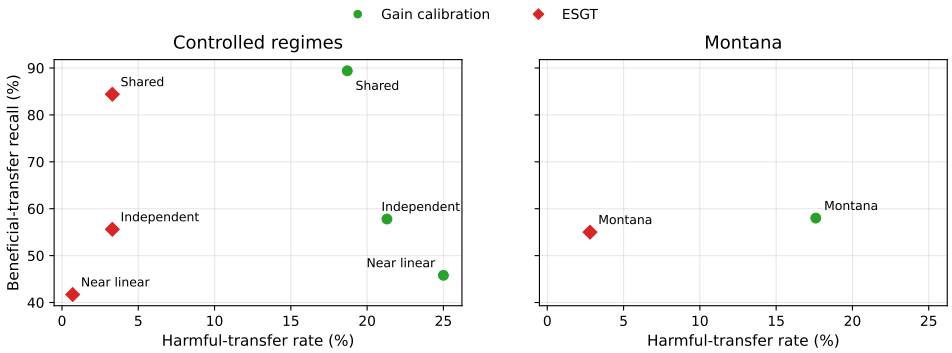


Figure 2: Transfer-decision quality. Lower harmful-transfer rate and higher beneficial-transfer recall are preferred. The left panel shows the three controlled regimes; the right panel shows Montana.

5.2 Controlled structural shift

Among the non-oracle methods, ESGT gives the lowest macro RMSE at 0.822, compared with 0.876 for shape-gated calibration and 0.884 for gain calibration. Its RMSE is 0.860 under shared structure and 1.035 under independent structure, 7.0% and 8.4% lower than the best alternative in those regimes. On near-linear targets, shape-gated calibration reaches 0.568°C and ESGT 0.570°C. The overall improvement therefore comes from the shared- and independent-structure regimes.

Under independent structure, ESGT reduces harm from 21.3% to 3.3% while retaining 55.6% of beneficial residual corrections, compared with 57.8% for gain calibration. Under shared structure, recall is 84.4% with 3.3% harm. Near-linear activation falls to 4.7%, where little residual headroom remains.

The three regimes stress different aspects of transfer. Independent structure combines useful source capacity with substantial mismatch, near-linear targets leave little headroom beyond the target trend, and shared structure contains more compatible source information. In the last case, the retention gate mainly removes the remaining harmful corrections.

5.3 Measured profile evaluation

We use Montana as the primary measured evaluation. In Table 3, ESGT reaches 0.370°C, compared with 0.403°C for PCHIP, 0.400°C for the selected fallback, and 0.399°C for shape-gated calibration. This corresponds to reductions of 8.2% relative to PCHIP and 7.3% relative to shape-gated calibration.

Table 2: Transfer-decision quality in percent. Harm compares the applied predictor with the target fallback in the same episode. Recall is conditional on a retrospectively beneficial unshrunk residual GP; Reject is conditional on a harmful candidate.

Regime	Rule	Active	Harm	Recall	Reject
Shared	Gain calibration	82.3	18.7	89.4	28.3
Shared	**ESGT**	54.7	**3.3**	84.4	**90.0**
Independent	Gain calibration	50.7	21.3	57.8	52.4
Independent	**ESGT**	21.3	**3.3**	55.6	**93.3**
Near linear	Gain calibration	40.3	25.0	45.8	60.1
Near linear	**ESGT**	4.7	**0.7**	41.7	**98.6**
Montana	Gain calibration	43.0	17.6	58.0	67.0
Montana	**ESGT**	25.6	**2.8**	55.0	**94.0**

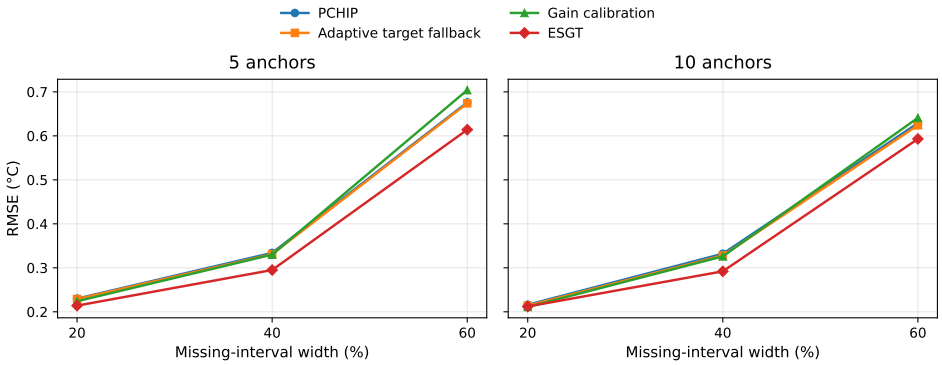


Figure 3: Montana RMSE by missing-interval width and anchor budget. The figure focuses on PCHIP, the adaptive target fallback, gain calibration, and ESGT.

Figure 3 breaks the Montana result down by gap width and anchor budget. ESGT gains most at 40% and 60% missingness, while gain calibration has a small advantage in one short-gap setting. On the report-derived field profiles, ESGT reaches 1.910°C, 4.3% below shape-gated calibration at 1.995°C.

Table 3: Profile reconstruction RMSE in °C. Montana uses equal-county aggregation over 30 wells in 17 counties. Reports denote report-derived field profiles and use equal-profile means.

Method	Montana	Reports
PCHIP	0.403	2.069
Target GP	0.769	2.857
Trend GP	0.528	3.082
Adaptive target fallback	0.400	2.025
Residual GP	0.793	3.539
RECaST, MAP adapter	1.244	3.525
TC_{MS}Stack, fixed kernel	0.858	2.832
Gain calibration	0.406	2.010
Shape-gated calibration	0.399	1.995
ESGT	**0.370**	**1.910**
Oracle coefficient, diagnostic	0.352	1.855

On Montana, ESGT activates in 25.6% of county-weighted episodes, lowers harm from 17.6% to 2.8%, and retains 55.0% of beneficial residual corrections, compared with 58.0% for gain calibration. Its paired MSE gains over the selected fallback and PCHIP are 0.0260 and 0.0273, using the same county weighting as the RMSE aggregation.

Differences are small at 20% missingness, while 40% and 60% gaps leave more unresolved structure for the source profile to improve.

5.4 SENSITIVITY AND STRESS TESTS

Geometry weighting and multi-anchor outer validation matter most under independent structure and on Montana. Removing geometry weighting raises RMSE from 1.035 to 1.107 under independent structure and from 0.370°C to 0.392°C on Montana; single-anchor outer blocks show a similar decline. Removing shape similarity also worsens the shared, independent, and Montana results. Shrinkage has a smaller effect, but using unit gain after retention is worse than ESGT in all four settings, so calibration remains useful even after the retention decision.

Table 4: Component sensitivity, mean RMSE in °C.

Variant	Shared	Independent	Near linear	Montana
ESGT	0.860	1.035	0.570	0.370
Gain calibration	0.925	1.150	0.577	0.406
Shape-gated calibration	0.930	1.130	0.568	0.399
No geometry weighting	0.910	1.107	0.573	0.392
Single-anchor outer blocks	0.921	1.114	0.574	0.395
No shrinkage	0.858	1.065	0.569	0.386
No shape similarity	0.878	1.085	0.569	0.391
Unit gain after retention	0.918	1.102	0.575	0.410

The stress tests vary source–target compatibility through shifted layer boundaries, a localized target anomaly, a mixed source pool with half of the sources mismatched, and a setting with only one admissible source. Appendix A gives the constructions.

ESGT improves on gain calibration under boundary shift, half-source mismatch, and one-source transfer by 8.3%, 9.7%, and 6.2%, respectively. Gain calibration is slightly better for the localized anomaly, with RMSEs of 2.642°C and 2.648°C. ESGT harm stays between 2.5% and 4.2% across the stress tests, indicating that broad mismatch is easier to detect from outer evidence than a highly localized anomaly. Appendix Table 5 gives the full results.

6 CONCLUSION

We have introduced ESGT for sparse geothermal profile reconstruction with limited target observations. The method separates correction calibration from source retention using geometry-matched pseudo-gaps, outer held-out evidence, and detrended shape agreement. When transfer is rejected, it returns to the selected target-only reconstruction.

In both the controlled and measured evaluations, ESGT reduced reconstruction error while limiting harmful transfer. Macro simulation RMSE is 0.822 versus 0.876 for the strongest non-oracle alternative, and Montana RMSE is 0.370°C versus 0.403°C for PCHIP. The remaining gap to the oracle separates into calibration shortfall and retention regret.

The current experiments use small target-anchor budgets, depth-span-based source matching, and a steady-state simulator. Broader field evaluation and source matching with richer geological and physical descriptors are natural next steps.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–

175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

David Blackwell, Maria Richards, and Cathy Chickering Pace. Nationwide heat flow, well data, and related files in the smu node of the national geothermal database system. Geothermal Data Repository, submission 1704; moved to GDR in 2026, 2014. URL https://gdr.openei.org/submissions/1704.

Erick R. Burns, Jacob DeAngelo, and Colin F. Williams. Updated three-dimensional temperature maps for the Great Basin, USA. In *Proceedings of the 49th Workshop on Geothermal Reservoir Engineering*. Stanford University, 2024. URL https://pangea.stanford.edu/ERE/db/GeoConf/papers/SGW/2024/Burns.pdf. SGP-TR-227.

Yuyang Deng and Samory Kpotufe. Mixed-sample SGD: An end-to-end analysis of supervised transfer learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3108. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/85d4d3ba2d10ca59a3cff87d5692474f-Abstract-Conference.html.

F. N. Fritsch and J. Butland. A method for constructing local monotone piecewise cubic interpolants. *SIAM Journal on Scientific and Statistical Computing*, 5(2):300–304, 1984. doi: 10.1137/0905021. URL https://doi.org/10.1137/0905021.

Pedro H. T. Gama, Jackson Faria, Jessica Sena, Francisco Neves, Vinícius R. Riffel, Lucas Perez, André Korenchendler, Matheus C. A. Sobreira, and Alexei M. C. Machado. Imputation in well log data: A benchmark for machine learning methods. *Computers & Geosciences*, 196:105789, 2025. doi: 10.1016/j.cageo.2024.105789. URL https://doi.org/10.1016/j.cageo.2024.105789.

Jonathan E. Gordon, Olatunde D. Akanbi, Deepa C. Bhuvanagiri, Hope E. Omodolor, Vibha Mandayam, Roger H. French, Jeffrey M. Yarus, and Erika I. Barcelos. Geospatial modeling of near subsurface temperatures of the contiguous united states for assessment of materials degradation. *Scientific Reports*, 15:1053, 2025. doi: 10.1038/s41598-024-85050-3. URL https://doi.org/10.1038/s41598-024-85050-3.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.01434.

Tomislav Hengl, Gerard B. M. Heuvelink, and David G. Rossiter. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315, 2007. doi: 10.1016/j.cageo.2007.05.001. URL https://doi.org/10.1016/j.cageo.2007.05.001.

Jimmy Hickey, Jonathan P. Williams, and Emily C. Hector. Transfer learning with uncertainty quantification: Random effect calibration of source to target (recast). *Journal of Machine Learning Research*, 25(338):1–40, 2024. URL https://jmlr.org/papers/v25/22-1369.html.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Multiple-source adaptation for regression problems, 2017. URL https://arxiv.org/abs/1711.05037. arXiv:1711.05037.

Xiaoyu Hu and Zhenhua Lin. Transfer learning meets functional linear regression: No negative transfer under posterior drift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17351–17358, 2025. doi: 10.1609/aaai.v39i16.33907. URL https://doi.org/10.1609/aaai.v39i16.33907.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804. URL https://doi.org/10.1016/j.patter.2023.100804.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A

benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Wouter M. Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, 2021. doi: 10.1109/TPAMI.2019.2945942. URL https://doi.org/10.1109/TPAMI.2019.2945942.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, pp. 1041–1048, 2008. URL https://papers.nips.cc/paper/2008/hash/0e65972dce68dad4d52d063967f0a705-Abstract.html.

Feisheng Mou, Yinhui Zuo, Junlong Zhang, Luquan Zhang, Guangzheng Jiang, Yingchun Wang, Jiaxin Xie, and Xinglong Wang. A borehole-geophysics workflow for constructing continuous thermal-property and temperature profiles from incomplete well logs: A case study from the sichuan basin. *Geothermics*, 141:103737, 2026. doi: 10.1016/j.geothermics.2026.103737. URL https://doi.org/10.1016/j.geothermics.2026.103737.

Ignavier Ng, Yan Li, Zijian Li, Yujia Zheng, Guangyi Chen, and Kun Zhang. A general representation-based approach to multi-source domain adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45911–45933, 2025. URL https://proceedings.mlr.press/v267/ng25a.html.

Daniel O'Malley, Christopher W. Johnson, Javier E. Santos, Pablo Lara, Sandro Malusà, Bharat Srikishan, John Kath, Arnab Mazumder, Mohamed Mehana, David Coblentz, Nathan DeBardeleben, Earl Lawrence, and Hari Viswanathan. In-context learning enables continental-scale subsurface temperature prediction from sparse local observations, 2026. URL https://arxiv.org/abs/2605.16665. arXiv preprint arXiv:2605.16665.

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. URL https://doi.org/10.1109/TKDE.2009.191.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881. URL https://doi.org/10.1111/ecog.02881.

Akihiro Shima, Kazuya Ishitsuka, Weiren Lin, Elvar K. Bjarkason, and Anna Suzuki. Modeling unobserved geothermal structures using a physics-informed neural network with transfer learning of prior knowledge. *Geothermal Energy*, 12:38, 2024. doi: 10.1186/s40517-024-00312-7. URL https://doi.org/10.1186/s40517-024-00312-7.

Hui Sun, Zheng Xie, Hao-Yuan He, and Ming Li. Mitigating negative transfer via reducing environmental disagreement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8485–8502, 2026. doi: 10.1109/TPAMI.2026.3672777. URL https://doi.org/10.1109/TPAMI.2026.3672777.

Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: A theory of transfer learning in high-dimensional regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58142–58168, 2025. URL https://proceedings.mlr.press/v267/tahir25a.html.

Yang Tan, Enming Zhang, Yang Li, Shao-Lun Huang, and Xiao-Ping Zhang. Transferability-guided cross-domain cross-task transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2423–2436, 2025. doi: 10.1109/TNNLS.2024.3358094. URL https://doi.org/10.1109/TNNLS.2024.3358094.

Pengfei Wei, Ramon Sagarna, Yiping Ke, Yew-Soon Ong, and Chi-Keong Goh. Source-target similarity modelings for multi-source transfer gaussian process regression. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3722–3731, 2017. URL https://proceedings.mlr.press/v70/wei17a.html.

Pengfei Wei, Thanh Vinh Vo, Xinghua Qu, Yew-Soon Ong, and Zejun Ma. Transfer kernel learning for multi-source transfer gaussian process regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3862–3876, 2023. doi: 10.1109/TPAMI.2022.3184696. URL https://doi.org/10.1109/TPAMI.2022.3184696.

Keru Wu, Yuansi Chen, Wooseok Ha, and Bin Yu. Prominent roles of conditionally invariant components in domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 26(110):1–92, 2025. URL https://jmlr.org/papers/v26/23-1234.html.

Yichen Xu, Ryumei Nakada, Linjun Zhang, and Lexin Li. Residual feature integration is sufficient to prevent negative transfer. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d998ff3e6ecce821986c9dd876be9bc-Abstract-Conference.html.

Kai Yang, Jie Lu, Wanggen Wan, and Guangquan Zhang. Multi-source transfer regression via source-target pairwise segment. *Information Sciences*, 556:389–403, 2021. doi: 10.1016/j.ins.2020.09.074. URL https://doi.org/10.1016/j.ins.2020.09.074.

Qingyue Zhang, Haohao Fu, Guanbo Huang, Yaoyuan Liang, Chang Chu, Tianren Peng, Yanru Wu, Qi Li, Yang Li, and Shao-Lun Huang. A high-dimensional statistical method for optimizing transfer quantities in multi-source transfer learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0858. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2459ff8e62f8be569d91e870139e16dd-Abstract-Conference.html.

Doudou Zhou, Mengyan Li, Yun Wang, Tianxi Cai, and Molei Liu. Domain adaptation targeting heterogeneous and imbalanced subgroups. *Journal of Machine Learning Research*, 27(181):1–65, 2026. URL https://www.jmlr.org/papers/v27/25-1766.html.

Wei Zhou, Haihang Zhao, Xiangchengzhen Li, Zhongli Qi, Fuqiang Lai, and Jun Yi. Missing well logs reconstruction based on cascaded bidirectional long short-term memory network. *Expert Systems with Applications*, 259:125270, 2025. doi: 10.1016/j.eswa.2024.125270. URL https://doi.org/10.1016/j.eswa.2024.125270.

A Detailed protocol

For each of the shared-structure, independent-structure, and near-linear regimes, validation uses 20 independently generated scenarios and final evaluation uses 50 separate scenarios. Each scenario contains four source wells and one target well, with gap widths 0.2, 0.4, and 0.6 and anchor budgets 5 and 10. The Montana evaluation contains 540 feasible episodes; ties in source-well ranking are resolved by well ID. The report-derived evaluation contains 10 feasible episodes.

Stress-test construction. For the boundary-shift condition, the internal target-layer boundaries are perturbed relative to the source structure while layer order and thermal properties are preserved. Each internal boundary is displaced by up to 20% of the smaller adjacent-layer thickness. The localized-anomaly condition adds a Gaussian-shaped temperature perturbation inside the missing interval,

$$\Delta T(x) = A \exp\left(-\frac{(x-c)^2}{2\sigma_a^2}\right), \tag{12}$$

where c lies inside the query interval, $|A| = 5°\text{C}$, and $\sigma_a = 0.05$ of the rescaled profile span. In the half-source-mismatch condition, two of the four source profiles share the target structure and the other two are generated independently. The one-source condition keeps only the highest-ranked source under the same depth-span criterion used in the main experiments.

Table 5: Controlled stress scenarios. RMSE values are in °C; ESGT harm is reported in percent.

Condition	Fallback	Gain calibration	ESGT	ESGT harm, %
Boundary shift	1.245	1.107	1.015	3.3
Localized anomaly	2.809	2.642	2.648	4.2
Half sources mismatched	1.214	1.096	0.990	2.5
One source	1.123	1.023	0.960	3.3

B Additional model and baseline details

Target-only GP candidates. Let $\mathcal{G}_A$ denote the Matérn-$3/2$ residual GP from the main text, with length scale 500 on the rescaled depth axis and nominal observation noise $0.5°\text{C}$. The Target GP uses the target-anchor mean $\bar{y}_A$ as its constant mean,

$$m_{\text{TGP},A}(x) = \bar{y}_A + \mathcal{G}_A\{y_A - \bar{y}_A\}(x). \tag{13}$$

The Trend GP first fits the affine temperature–depth trend $\ell_A(x) = \beta_0 + \beta_1 x$ and then applies the same GP to the residuals,

$$m_{\text{TrGP},A}(x) = \ell_A(x) + \mathcal{G}_A\{y_A - \ell_A(A)\}(x). \tag{14}$$

Outer-block geometry weighting. For outer block H, let h_H be the span between the retained anchors that bracket it. Its unnormalized geometry weight is

$$\widetilde{w}_H = \exp\left(-\frac{|\log(h_H/h_Q)|}{\tau}\right), \qquad w_H = \frac{\widetilde{w}_H}{\sum_{H'} \widetilde{w}_{H'}}, \tag{15}$$

with the same $\tau = 0.5$ used for inner calibration blocks.

Calibration baselines. Gain calibration uses the same adaptive fallback, source candidate, and geometry-matched coefficient as ESGT, but omits the outer evidence gate,

$$\hat{y}_A^{\text{gain}}(x) = b_A(x) + \hat{g}_A(h_Q)[t_A(x) - b_A(x)]. \tag{16}$$

Shape-gated calibration adds the requirement that the detrended shape statistic exceed the selected threshold,

$$\hat{y}_A^{\text{shape}}(x) = b_A(x) + \mathbf{1}\{s_A \geq \theta_s\}\hat{g}_A(h_Q)[t_A(x) - b_A(x)]. \tag{17}$$

C HYPERPARAMETER SELECTION AND SENSITIVITY

Hyperparameters were selected on the simulated validation split, which contains 20 independently generated scenarios from each of the shared-structure, independent-structure, and near-linear regimes. The final 150 simulation scenarios, Montana, and the report-derived profiles were reserved for evaluation.

Selection used two stages over coarse candidate grids. We first chose the geometry temperature τ and shrinkage coefficient ρ by equal-regime validation RMSE for the calibrated source-assisted predictor. With those values fixed, we selected the denominator floor δ_{floor} and the gate thresholds $(\theta_S, \theta_R, \theta_s)$ jointly. Configurations with validation harm above 5% were excluded; among the remainder, we chose the largest equal-regime relative MSE gain, breaking ties by higher beneficial-transfer recall and then lower activation. The selected configuration was then used unchanged in all evaluations.

Table 6: Hyperparameter search spaces and the selected ESGT configuration. All parameters were selected using only the simulated validation split.

Parameter	Role	Candidate values	Selected
τ	Geometry matching	$\{0.25, 0.50, 1.00\}$	0.50
ρ	Correction shrinkage	$\{0, 0.10, 0.20, 0.40\}$	0.20
δ_{floor}	Relative-gain denominator floor	$\{0.10, 0.25, 0.50\}$	0.25
θ_S	Relative-gain threshold	$\{0, 0.025, 0.05, 0.10, 0.20\}$	0.05
θ_R	Cross-block consistency threshold	$\{0.40, 0.50, 0.60, 0.75\}$	0.50
θ_s	Shape-similarity threshold	$\{0.55, 0.65, 0.70, 0.75, 0.85\}$	0.70

Larger values of τ allow looser geometry matches. The shrinkage grid includes the unshrunk case $\rho = 0$, and the gate grids cover both permissive and conservative settings.

Table 7: Local sensitivity to geometry matching, shrinkage, and the relative-gain denominator floor. One parameter is varied at a time while all other parameters remain at their selected values. Harm is reported in percent.

Parameter	Value	Sim. RMSE	Sim. harm	Montana RMSE	Montana harm
τ	0.25	0.829	2.0	0.368	2.5
	0.50	0.822	2.4	0.370	2.8
	1.00	0.835	3.4	0.381	3.6
ρ	0	0.831	3.5	0.386	4.1
	0.10	0.824	2.9	0.369	3.2
	0.20	0.822	2.4	0.370	2.8
	0.40	0.828	1.8	0.377	2.1
δ_{floor}	0.10	0.824	3.2	0.368	3.7
	0.25	0.822	2.4	0.370	2.8
	0.50	0.829	1.7	0.378	2.0

Table 7 varies one calibration parameter at a time around the selected configuration. At $\tau = 0.25$, Montana RMSE is 0.368 while simulation macro RMSE rises to 0.829; at $\tau = 1.00$, both results worsen. Removing shrinkage increases harm and RMSE under structural mismatch, whereas stronger shrinkage lowers harm at a modest RMSE cost. The denominator floor shows the same trade-off: smaller values admit more corrections, while larger values reject more of them.

Table 8 shows the local trade-off around the selected gate thresholds. Relaxing a threshold admits more source corrections and generally increases harm; tightening it reduces harm but can raise RMSE by rejecting useful corrections. The selected values give a middle operating point between these effects.

Table 8: Local sensitivity to the three ESGT gate thresholds. One threshold is varied at a time while the remaining parameters are fixed at their selected values. Harm is reported in percent.

Threshold	Value	Sim. RMSE	Sim. harm	Montana RMSE	Montana harm
θ_S	0.025	0.823	3.8	0.368	4.2
	0.050	0.822	2.4	0.370	2.8
	0.100	0.828	1.5	0.377	1.8
θ_R	0.40	0.824	3.5	0.369	3.9
	0.50	0.822	2.4	0.370	2.8
	0.60	0.829	1.7	0.379	2.0
θ_s	0.65	0.825	3.6	0.368	4.0
	0.70	0.822	2.4	0.370	2.8
	0.75	0.828	1.8	0.378	2.1

D Matched literature baseline implementations

D.1 TC$_{MS}$Stack fixed-kernel adapter

TC$_{MS}$Stack (Wei et al., 2017) combines single-source transfer Gaussian processes with weights determined by source-target similarity coefficients. For source i, $\lambda_i \in [-0.995, 0.995]$ is the transfer coefficient. Temperatures are centered and scaled using the target anchors. We use a squared-exponential kernel with length scale 500 on the rescaled depth axis and nominal observation noise $0.5°$C. The similarity coefficients are optimized jointly against target-anchor leave-one-out squared error using limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization with box constraints (L-BFGS-B) and deterministic starts. We refer to this setup as the fixed-kernel adapter.

D.2 RECaST MAP adapter

For continuous outcomes, RECaST (Hickey et al., 2024) recalibrates the source prediction as

$$y_j = \beta_j f_s(a_j) + \epsilon_j, \qquad \beta_j \sim \text{Cauchy}(\delta, \gamma), \qquad \epsilon_j \sim \mathcal{N}(0, \sigma^2). \tag{18}$$

Here β_j is the random calibration effect for target anchor j, δ and γ are the Cauchy location and scale, and ϵ_j is Gaussian observation noise with standard deviation σ. The Gaussian-Cauchy convolution is evaluated with the Voigt density. Priors are $\delta \sim \mathcal{N}(1, 100)$, $\log\gamma \sim \mathcal{N}(-2, 4)$, and $\log\sigma \sim \mathcal{N}(0, 4)$. MAP optimization uses L-BFGS-B. The point prediction is $\hat{\delta} f_s(x)$.

E Diagnostic source-value accounting

The following quantities are computed from query temperatures after prediction and are used only for retrospective analysis. Let b_A be the selected target fallback, t_A the source-assisted candidate, and $d_A = t_A - b_A$. For a mixing coefficient $g \in [0, 1]$,

$$\hat{y}_g = b_A + g d_A. \tag{19}$$

With query error $e_Q = y_Q - b_A(Q)$ and query-set correction $d_Q = d_A(Q)$, define

$$C_Q = \langle e_Q, d_Q \rangle_Q, \qquad D_Q = \langle d_Q, d_Q \rangle_Q. \tag{20}$$

The MSE gain over the fallback is

$$G_Q(g) = \text{MSE}_Q(b_A) - \text{MSE}_Q(\hat{y}_g) = 2gC_Q - g^2 D_Q. \tag{21}$$

For $D_Q > 0$, the oracle coefficient is $g_Q^* = \text{clip}_{[0,1]}(C_Q/D_Q)$; otherwise $g_Q^* = 0$. The oracle headroom is $H_Q = G_Q(g_Q^*)$. Let $\hat{g}$ be the calibrated coefficient before retention, let $r_A \in \{0, 1\}$ be the final gate decision, and define $V_Q = G_Q(\hat{g})$. The calibration shortfall and retention regret are

$$K_Q = H_Q - \max(V_Q, 0), \qquad L_Q = \max(V_Q, 0) - r_A V_Q. \tag{22}$$

The realized source gain can therefore be written as

$$G_Q(r_A \hat{g}) = H_Q - K_Q - L_Q. \tag{23}$$

Thus realized gain equals oracle headroom minus calibration shortfall and retention regret, as summarized in the main text.

Expanding the squared query error gives

$$\|e_Q - g d_Q\|_Q^2 = \|e_Q\|_Q^2 - 2gC_Q + g^2 D_Q, \tag{24}$$

which gives $G_Q(g)$. Since g_Q^* maximizes this gain over $[0, 1]$, $H_Q \geq G_Q(\hat{g})$ and $K_Q \geq 0$. The term L_Q captures the difference between the best decision available for the calibrated candidate and the decision made by the retention gate.

Table 9: MSE-scale source-value accounting. Realized gain equals headroom minus calibration shortfall minus retention regret.

Regime	Headroom	Calibration	Retention	Gain
Shared	1.4200	0.3200	0.2700	0.8300
Independent	0.8200	0.2900	0.1600	0.3700
Near linear	0.0660	0.0190	0.0430	0.0040
Montana	0.0500	0.0100	0.0140	0.0260

F FINITE-ANCHOR AMBIGUITY

Fix the observed anchors, source information, target fallback b_A, and a nonzero correction v_Q on the query set. Consider two targets that agree at every observed anchor but take query values $b_A(Q)+v_Q$ and $b_A(Q) - v_Q$. A smooth interpolant can pass through either finite assignment, so the retention rule receives the same observed inputs in both cases. For the first target, adding v_Q removes the fallback error $\|v_Q\|_Q^2$. For the second, the same correction changes a fallback error $\|v_Q\|_Q^2$ into $4\|v_Q\|_Q^2$, an excess MSE of $3\|v_Q\|_Q^2$. The same observed anchors can therefore support opposite transfer outcomes unless additional structure is assumed inside the gap.

G PRACTICAL INTERPRETATION

Selective transfer is most useful when the source explains structure that the target anchors do not resolve. The selected target-only model supplies the fallback, and nested evaluation removes many harmful residual corrections while keeping useful ones. Near-linear targets, one short-gap field setting, the shared-structure no-shrinkage variant, and the localized-anomaly stress test are cases in which a simpler alternative performs slightly better.

H EXPANDED RESULTS

Table 10 reports Montana errors by gap width and anchor budget. ESGT has its largest advantage at 40% and 60% missingness, where the target-only reconstruction leaves more structure unresolved. At 20% missingness, the methods are close, and Gain is slightly better than ESGT with 10 anchors. This agrees with the main result that source retention matters most when the target anchors provide weaker constraints.

Table 10: Montana RMSE in °C by gap width and anchor budget.

Gap	Anchors	PCHIP	Fallback	Gain	ESGT
20%	5	0.230	0.229	0.224	0.214
20%	10	0.216	0.214	0.211	0.212
40%	5	0.334	0.331	0.330	0.295
40%	10	0.332	0.328	0.326	0.292
60%	5	0.676	0.674	0.704	0.614
60%	10	0.629	0.624	0.641	0.593

Table 11 reports paired MSE differences between ESGT and selected comparators, with positive values indicating lower error for ESGT. The largest gain appears in the shared-structure regime, while the near-linear regime shows little difference because the target-only reconstruction is already strong. On Montana, the paired gains remain positive against both the adaptive fallback and PCHIP.

Table 11: Paired MSE gain, comparator minus ESGT. Positive values favor ESGT.

Dataset or regime	Comparator	Units	Gain
Shared	Fallback	50	0.8300
Independent	Fallback	50	0.3700
Near linear	Fallback	50	0.0040
Montana	Fallback	17	0.0260
Montana	PCHIP	17	0.0273

I AI USE STATEMENT

We used large language models as writing assistants to improve the concision and clarity of the manuscript and to express our ideas more precisely. The scientific content, technical contributions, and conclusions of this work were developed independently by the authors. The authors take full responsibility for all ideas, results, and claims presented in this paper.