THE KNOWABILITY GAP: AUDITING WHEN BETTER MODELS CANNOT REPLACE MISSING INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent prediction error can have two different causes: the learner may be inadequate, or the observation may omit information needed to determine the target. We study how to distinguish these cases before spending more computation on model scaling. For an observed input, target, declared environment family, and task metric, we define the *Knowability Gap* (K-Gap) as the Wasserstein Chebyshev radius of the compatible environment-conditioned target laws. Under the rectangular envelope that allows any locally admissible mechanism, K-Gap equals the optimal robust Bayes risk; for globally coupled mechanism classes contained in that envelope, it is a conservative upper audit rather than an assumption-free deployment limit. Our main finite-environment theorem gives a two-sided decomposition that separates conditional-law estimation error from environment-coverage error, and reduces to an observed-environment certificate when the declared family is finite. We also characterize when side information is guaranteed to reduce the radius and show why arbitrary conditioning need not do so. Controlled experiments recover an exact distributional K-Gap of 0.575 and verify the sample/coverage correction over 4,500 trials. On real hyperspectral data, a five-site audit is inconclusive under a frozen support rule. On GLORIA, extending 400–600 nm with 602–800 nm lowers test-size-weighted Chl-a MAE from 21.46 to 15.60; the estimated Mean K-Gap point estimate also decreases, although its campaign-bootstrap interval crosses zero. The real extension is more favorable than all 20 frozen randomized-information controls, and across four model families Medium $X + A$ has lower mean MAE than Large X-only. The resulting audit separates four often-conflated resources: model capacity, repeated samples, environment coverage, and measurement information.

1 INTRODUCTION

When prediction error stops improving, the usual response is to scale the learner: use a larger model, collect more labels, or optimize longer. That response is appropriate when the learner is the bottleneck. It cannot recover information that is absent from every input the learner receives.

This distinction matters in scientific prediction. A sensor may map physically different states to nearly the same observation; a clinical measurement may omit the variable that separates two outcomes; a remote-sensing band may not resolve the mechanism that drives the target. In each case, more capacity can fit the observed view better without making the missing variable observable. The practical question should therefore come before model selection:

> *Before scaling a predictor, can we tell whether the bottleneck is the learner or the measurement—and which additional resource would actually reduce the best achievable risk?*

Existing tools answer neighboring questions. Predictive uncertainty quantifies uncertainty of a fitted model (Lakshminarayanan et al., 2017; Ovadia et al., 2019); OOD detection asks whether an input looks atypical (Liu et al., 2020); conformal methods provide coverage under explicit exchangeability or shift conditions (Angelopoulos and Bates, 2023; Barber et al., 2023); domain adaptation, invariant learning, and group robustness seek predictors that transfer across distributions (Ben-David et al.,

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

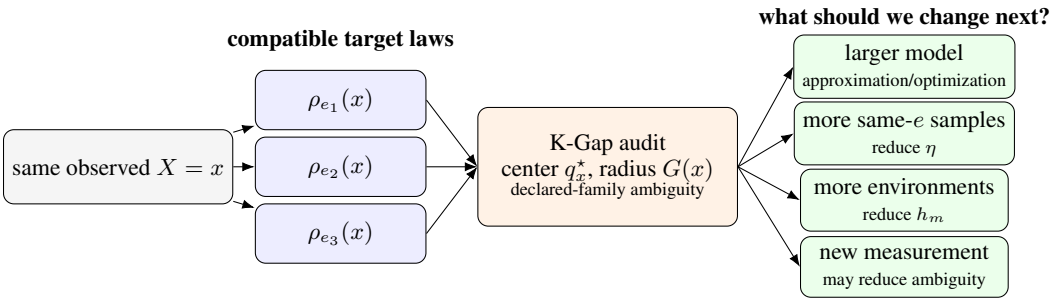

Figure 1: K-Gap is an audit of a declared prediction problem. Several target laws can remain compatible with the same observation. The theory separates uncertainty due to finite estimation and finite environment coverage from ambiguity attached to the information itself. The four remedies therefore act on different bottlenecks.

2010; Courty et al., 2017; Arjovsky et al., 2019; Sagawa et al., 2020; Gulrajani and Lopez-Paz, 2021; Koh et al., 2021). None of these quantities, by itself, asks how much disagreement remains among target mechanisms that are still compatible with the same observation.

Inverse problems provide the closest classical analogy: injectivity and stability ask whether latent states are recoverable (Engl et al., 1996; Stuart, 2010), while partial identification describes compatible sets when point identification fails (Manski, 2003). Our goal is different from reconstructing the entire latent world. We ask for a task-level audit: among the target laws still compatible with what the predictor sees, what common prediction is minimax, how large is the remaining radius, and what kind of additional data can shrink uncertainty about that radius?

For $X = x$, let $\rho_e(x) = P_e(Y \in \cdot \mid X = x)$ denote the target law in environment e. We define the *Knowability Gap* (K-Gap) as the smallest Wasserstein radius covering the declared compatible laws. The abstract circumradius is classical; our contribution is the conditional prediction formulation and its audit consequences. The rectangular class used for the exact Bayes identity is deliberately conservative. We therefore prove a sandwich showing how this envelope relates to globally coupled mechanism classes rather than treating rectangularity as an implicit fact about deployment. Our main finite-data result then separates two errors that model scaling cannot distinguish: inaccurate estimation of mechanisms already observed and incomplete coverage of the admissible mechanism family.

Contributions.

- **A conservative task-level audit with an explicit coupling boundary.** We formulate K-Gap as a conditional Wasserstein Chebyshev radius. It equals robust Bayes risk for the rectangular envelope, and upper-bounds the robust risk of any globally coupled mechanism class contained in that envelope. This states exactly when the radius is an attainable robust limit and when it should instead be read as a conservative audit.

- **A finite-data decomposition that separates two acquisition problems.** Our main theorem bounds population K-Gap by an empirical radius plus conditional-law estimation error η and environment-coverage error Lh_m. When the declared family is exactly the observed finite environments, $h_m = 0$, giving an observed-environment certificate without an extrapolation metric. Repeated samples and additional environments therefore reduce different uncertainties.

- **A failure-aware empirical audit.** Controlled tests recover an exact distributional K-Gap and verify the sample/coverage correction. Real-data experiments deliberately separate certified and non-certified evidence: Five-site wastewater is inconclusive under the frozen overlap diagnostic; GLORIA is a mean-level measurement audit with campaign uncertainty, randomized-information controls, and a four-model capacity comparison. Negative metadata and estimator-sensitivity results are retained rather than filtered out.

2 WHAT IS THE INFORMATION LIMIT AT FIXED OBSERVATION?

Let $E \in \mathcal{E}$ be an environment, $X \in \mathcal{X}$ the available observation, and $Y \in \mathcal{Y}$ the target. For each e, let $\rho_e(x) = P_e(Y \in \cdot \mid X = x)$ be a regular conditional law. Target laws lie in $(\mathcal{P}_p(\mathcal{Y}), \mathsf{W}_p)$, where the ground metric represents task cost. We compare environments only on a declared overlap measure ν supported where they are jointly evaluable; the overlap rule is frozen before target-test labels are inspected.

Definition 1 (Pointwise and integrated K-Gap)**.** For a nonempty admissible set $\mathcal{E}_x \subseteq \mathcal{E}$, define

$$\mathsf{G}(x) = \inf_{q \in \mathcal{P}_p(\mathcal{Y})} \sup_{e \in \mathcal{E}_x} \mathsf{W}_p(q, \rho_e(x)), \tag{1}$$

$$\mathsf{G} = \int \mathsf{G}(x)\, d\nu(x). \tag{2}$$

When the infimum in eq. (1) is attained, any minimizer is a Wasserstein Chebyshev center and is denoted $q_x^\star$; otherwise we work with measurable ε-centers. K-Gap is task-, environment-, support-, and metric-conditional, not an absolute property of Y.

For small scalar problems we also use the mean specialization

$$\mathsf{G}_{\mathrm{mean}}(x) = \frac{1}{2}\left(\sup_{e \in \mathcal{E}_x} \mu_e(x) - \inf_{e \in \mathcal{E}_x} \mu_e(x)\right), \qquad \mu_e(x) = \mathbb{E}_e[Y \mid X = x]. \tag{3}$$

It tests stability of the conditional mean; the distributional definition is needed when environments differ in tails, multimodality, or mass displacement.

Why Wasserstein, and what is metric-generic? The minimax-radius construction and robust-risk identity below are not tied to one distance. Wasserstein is useful here because the target ground metric can encode downstream cost, W_1 has a direct decision interpretation, and one-dimensional centers are computationally accessible through quantiles. Kantorovich–Rubinstein duality gives

$$\mathrm{diam}_{\mathsf{W}_1}\, \mathcal{S}(x) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left[\sup_{\rho \in \mathcal{S}(x)} \mathbb{E}_\rho f - \inf_{\rho \in \mathcal{S}(x)} \mathbb{E}_\rho f\right], \tag{4}$$

where $\mathcal{S}(x) = \{\rho_e(x) : e \in \mathcal{E}_x\}$. Thus the W_1 diameter is the largest disagreement over one-Lipschitz downstream utilities and $\frac{1}{2}\operatorname{diam}\mathcal{S}(x) \le \mathsf{G}(x) \le \operatorname{diam}\mathcal{S}(x)$. A Wasserstein barycenter minimizes an *average* transport cost (Agueh and Carlier, 2011); the K-Gap center minimizes the *worst-case* distance and returns a minimax prediction. Pairwise diameter returns only a worst pair and no common prediction.

3 WHAT ROBUST RISK DOES K-GAP REPRESENT?

3.1 RECTANGULAR AMBIGUITY IS AN ENVELOPE, NOT A HIDDEN DEPLOYMENT ASSUMPTION

Let $\mathcal{S}(x) = \{\rho_e(x) : e \in \mathcal{E}_x\}$. The rectangular envelope $\mathrm{Sel}(\mathcal{S})$ contains all measurable selections $x \mapsto \rho_x \in \mathcal{S}(x)$, so the active mechanism may vary with x. This is the largest locally compatible class and therefore a conservative audit object.

Theorem 2 (Robust Bayes identity)**.** *Assume $\mathcal{S}$ has nonempty compact values and a measurable graph, and that $\int \sup_{\rho \in \mathcal{S}(x)} \mathsf{W}_p(\rho, \delta_{y_0}) d\nu(x) < \infty$ for some y_0. Then*

$$\inf_{q : \mathcal{X} \to \mathcal{P}_p(\mathcal{Y})} \sup_{\rho(\cdot) \in \mathrm{Sel}(\mathcal{S})} \int \mathsf{W}_p(q_x, \rho_x) d\nu(x) = \int \mathsf{G}(x) d\nu(x) = \mathsf{G}. \tag{5}$$

Measurable ε-centers attain the value up to ε; uniquely attained measurable centers attain it exactly.

The identity gives the radius a prediction meaning, but only for the declared rectangular envelope. Real environments can obey global coupling constraints, for example one site or physical regime remaining fixed across all x. The next result makes that conservatism explicit.

Proposition 3 (Rectangular-envelope sandwich). *Let $\mathfrak{M} \subseteq \mathrm{Sel}(\mathcal{S})$ be any measurable globally coupled mechanism class and define*

$$R^\star(\mathfrak{M}) = \inf_q \sup_{m\in\mathfrak{M}} \int \mathsf{W}_p(q_x, \rho_x^m) d\nu(x).$$

Then $R^\star(\mathfrak{M}) \le \mathsf{G}$. If $\mathfrak{M}$ contains every fixed-environment mechanism $m_e : x \mapsto \rho_e(x)$, then

$$\inf_q \sup_{e\in\mathcal{E}} \int \mathsf{W}_p(q_x, \rho_e(x)) d\nu(x) \;\le\; R^\star(\mathfrak{M}) \;\le\; \mathsf{G}. \tag{6}$$

Both inequalities can be strict; equality on the right holds for the full rectangular envelope.

Thus K-Gap is exact for local rectangular uncertainty and conservative for any tighter globally coupled class. If scientific knowledge specifies a smaller class $\mathfrak{M}$, the corresponding robust risk should replace the envelope rather than being hidden inside the word "deployment."

3.2 What model scaling can and cannot change

Corollary 4 (Asymptotic scaling within the declared envelope). *Let $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$ be conditional distribution predictors whose union is dense in $L^1(\nu; \mathcal{P}_p(\mathcal{Y}))$ under W_p. If $R_k^\star$ is the optimal rectangular robust risk over $\mathcal{H}_k$, then $R_k^\star \downarrow \mathsf{G}$. If, for each fixed k, an empirical learner is risk-consistent in $\mathcal{H}_k$ as within-environment sample size $n \to \infty$, then*

$$\lim_{k\to\infty} \lim_{n\to\infty} \widehat{R}_{k,n} = \mathsf{G}. \tag{7}$$

A joint sequence $k = k(n)$ requires uniform consistency or a controlled growth condition.

The statement is asymptotic and conditional on the declared envelope. A finite empirical plateau is not, by itself, evidence of an information limit.

4 Finite data: more samples are not more environments

Let $(\mathcal{E}, d_\mathcal{E})$ be compact and suppose the conditional mechanism varies smoothly across environments.

Assumption 5 (Environment regularity). For ν-a.e. x, $\mathsf{W}_p(\rho_e(x), \rho_{e'}(x)) \le L(x) d_\mathcal{E}(e, e')$ for all $e, e' \in \mathcal{E}_x$, with $L \in L^1(\nu)$.

For observed environments $e_1, \ldots, e_m$ and conditional-law estimates $\widehat{\rho}_j(x)$, define

$$\widehat{\mathsf{G}}_m(x) = \inf_q \max_j \mathsf{W}_p(q, \widehat{\rho}_j(x)), \qquad \eta(x) = \max_j \mathsf{W}_p(\widehat{\rho}_j(x), \rho_{e_j}(x)),$$
$$h_m(x) = \sup_{e\in\mathcal{E}_x} \min_j d_\mathcal{E}(e, e_j).$$

Theorem 6 (Environment-and-sample decomposition). *For ν-a.e. x,*

$$[\widehat{\mathsf{G}}_m(x) - \eta(x)]_+ \le \mathsf{G}(x) \le \widehat{\mathsf{G}}_m(x) + \eta(x) + L(x) h_m(x). \tag{8}$$

The same inequality integrates over ν.

This is the main finite-data result. The two errors answer different acquisition questions. More samples from the same environments can shrink η; they cannot shrink h_m. More environments can shrink h_m; they do not change the underlying population radius. The theorem therefore distinguishes "measure the same mechanisms more accurately" from "cover more mechanisms."

Corollary 7 (Observed-environment certificate). *If the declared admissible family at x is exactly the observed finite set $\{e_1, \ldots, e_m\}$, then $h_m(x) = 0$ and*

$$[\widehat{\mathsf{G}}_m(x) - \eta(x)]_+ \le \mathsf{G}_{\mathrm{obs}}(x) \le \widehat{\mathsf{G}}_m(x) + \eta(x). \tag{9}$$

Thus observed-family certification needs calibrated conditional-law error but no environment metric or extrapolation constant.

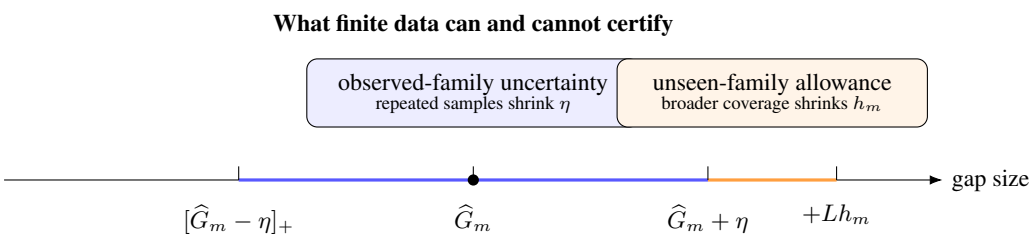

Figure 2: The certificate separates estimation from extrapolation. The blue interval concerns the declared observed environments. Extending the statement to unseen admissible environments adds the one-sided Lh_m term and therefore requires an environment metric and regularity assumption.

For scalar $Y \in [a, b]$, one concrete calibration route uses fixed representation cells and an independent calibration split. If the calibration targets are conditionally i.i.d. within each fixed environment–cell pair and n_{jb} samples from environment j fall in cell b, the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (Dvoretzky et al., 1956; Massart, 1990) gives

$$\eta_{jb}^{(1)} = (b - a)\sqrt{\frac{\log(2mB/\delta)}{2n_{jb}}} \tag{10}$$

for W_1, simultaneously over populated pairs. This calibration formula is used only where its sampling assumptions are justified; correlated field campaigns are not declared certified by applying DKW mechanically.

5 What can new information change?

A new variable can help in two different ways. It may identify which already-compatible mechanism is active, or it may change the within-environment conditional law itself. These cases should not be conflated.

Proposition 8 (Safe mechanism-revealing side information)**.** *Let A be available at prediction time. For ν-a.e. x and $\kappa(da \mid x)$-a.e. a, suppose (i) $\mathcal{E}_{x,a} \subseteq \mathcal{E}_x$ and (ii) A is mechanism-preserving,*

$$P_e(Y \in \cdot \mid X = x, A = a) = \rho_e(x), \qquad e \in \mathcal{E}_{x,a}. \tag{11}$$

Then the refined radius $\mathsf{G}_A(x, a) = \inf_q \sup_{e \in \mathcal{E}_{x,a}} \mathsf{W}_p(q, \rho_e(x))$ satisfies $\mathsf{G}_A(x, a) \le \mathsf{G}(x)$ pointwise and after integration.

The condition is substantive, not automatic. Without it, conditioning can reveal disagreement that was averaged out: if X is constant, $A \sim \mathrm{Bernoulli}(1/2)$, $Y = A$ in one environment and $Y = 1 - A$ in another, then $\mathsf{G}_X = 0$ but $\mathsf{G}_{X,A} = 1/2$ under W_1 on $\{0, 1\}$. Extra spectral bands in GLORIA are therefore treated as an empirical measurement extension, not as a verification of Proposition 8.

Proposition 9 (Deployment identifiability boundary)**.** *No finite-sample algorithm can uniformly estimate a universal deployment K-Gap without restrictions linking observed and deployment mechanisms: two meta-worlds can induce the same observed training distribution and different deployment K-Gaps.*

This boundary explains why the environment family, coupling model, support rule, and target metric are part of the audit statement. K-Gap is not presented as an assumption-free scalar attached to a dataset.

6 From theory to an audit

We distinguish three evidence levels. A *population K-Gap* is defined by the true environment-conditioned laws. An *observed-family certificate* additionally requires valid calibration of η in Corollary 7. An *estimated Mean K-Gap* is a weaker model-based proxy for conditional-mean disagreement. These names are not interchangeable.

The audit is ordered to make unsupported claims fail early. We first declare the environment family, coupling convention, target metric, and overlap rule without target-test labels. We next check whether environments are jointly evaluable on the chosen support. Only then do we estimate a gap and report it together with held-out risk and its evidence level. A candidate measurement is evaluated on matched samples and compared with randomized information. If support or calibration is insufficient, the output is *inconclusive* rather than a numerical deployment claim.

For scalar certified settings, fixed cells and an independent calibration split provide a direct route to Corollary 7. For high-dimensional or correlated field data where uniform conditional-law calibration is not established, we deliberately stop at the mean-level proxy. This is the status of the GLORIA experiment below; the controlled study is where the distributional quantity and finite-sample correction are tested directly.

7 CONTROLLED TESTS: DO THE THEORETICAL DISTINCTIONS APPEAR IN DATA?

The controlled studies are unit tests: the population answer is fixed before fitting, so the experiments check whether estimation, scaling, coverage, and added information behave as the theory predicts.

Exact distributional radius. Let $X \sim \mathrm{Unif}(0,1)$, $E \in \{-1,+1\}$ be unobserved, and $Y = g(X) + ER(X) + \epsilon$ with $R(X) = 0.15 + 0.85X$. At fixed x, the two conditional laws are translations by $2R(x)$; under W_1 their distance is $2R(x)$ and the midpoint translated law has radius $R(x)$. Hence the full distributional K-Gap is exactly $\mathsf{G}(x) = R(x)$ and $\mathsf{G} = \mathbb{E}R(X) = 0.575$. Increasing model capacity or same-view sample size approaches this value from above; the largest tested model reaches robust risk 0.592. Revealing E changes the information set and reduces risk to 0.001.

Sample error, coverage, and distributional ambiguity. With continuous $E \in [-1,1]$, an observed-environment radius can miss unobserved mechanisms. Adding the Lh_m term in Theorem 6 covers the population gap in all $4{,}500$ simulated trials. A noisy mechanism-revealing observation $A = E + \xi$ lowers residual ambiguity and prediction error as the sensor improves, matching Proposition 8. A separate construction shows why the full distributional object is richer than the mean specialization: two target laws share the same mean and variance but have $W_1 = 0.911$, $W_2 = 0.919$, and a one-third difference in $P(|Y| > 1.2)$.

8 REAL-DATA AUDIT: ABSTENTION, MEASUREMENT EXTENSION, AND CAPACITY

Real data answer a narrower question than the controlled tests. We do not claim a certified deployment K-Gap. We ask whether the audit rejects unsupported comparisons, whether a fixed measurement extension improves held-out prediction under matched samples, and whether the mean-gap proxy moves in the same direction under a frozen estimator.

8.1 FIVE-SITE WASTEWATER: WHEN THE AUDIT SHOULD STOP

The five-site wastewater benchmark contains 505 paired hyperspectral samples for turbidity, DOC, and NH_4–N on a shared 400–798 nm range (Lechevallier et al., 2025). We first reproduced the published PLS trends, then applied the support diagnostic frozen for the audit. Site identity is predictable from the spectrum (balanced accuracy 0.89 for logistic regression and 0.79 for random forest), so overlap cannot be assumed from domain similarity alone. Under the fixed-cell construction, no cell has enough calibration samples from all five sites; a direct 200-dimensional spectrum sensitivity check reaches the same decision. We therefore report *inconclusive under this support protocol*. This does not assert that mathematical overlap is absent under every possible diagnostic.

8.2 GLORIA: A MATCHED MEASUREMENT-EXTENSION TEST

GLORIA provides globally collected hyperspectral reflectance with water-quality targets (Lehmann et al., 2023). We freeze two sensor panels and define environments as campaigns within sensor. The

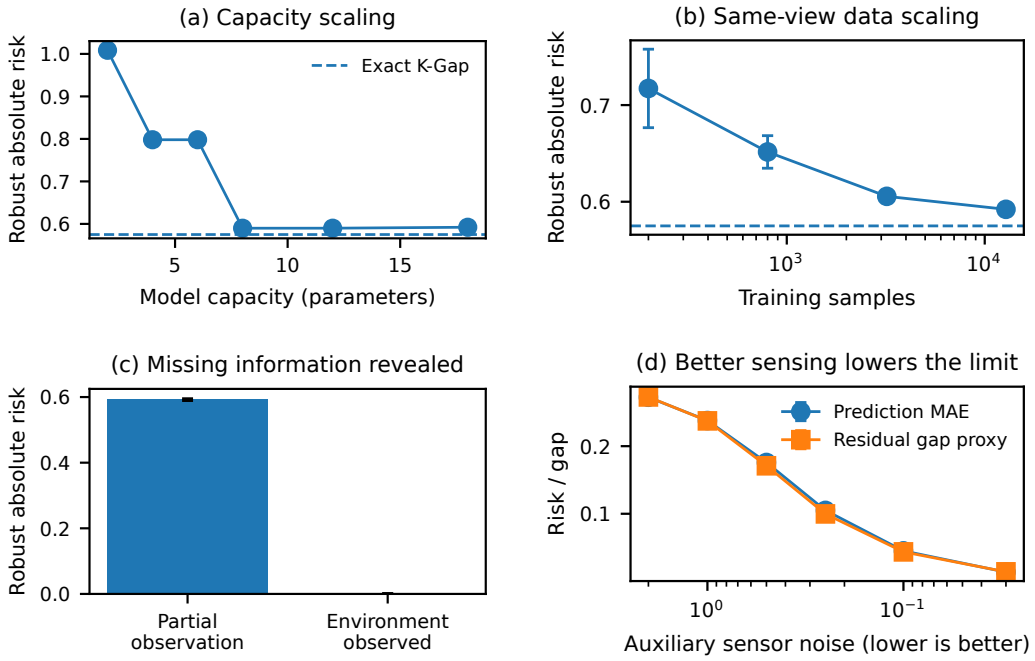

Figure 3: Controlled verification. Capacity and same-view data approach the known 0.575 radius; revealing the hidden mechanism changes the attainable risk; and a more accurate mechanism-revealing sensor lowers both residual ambiguity and prediction error. These panels verify implementation behavior, not a new empirical law.

Table 1: Primary GLORIA Chl-a/RF measurement audit. Brackets are 95% campaign-cluster bootstrap intervals for the paired change; the final column ranks the real extension against 20 frozen randomized-A controls. $\widehat{G}_{\text{mean}}$ is an estimated mean-level proxy, not a certified population gap.

Aggregation	MAE: $X \to X{+}A$	$\widehat{G}_{\text{mean}}$: $X \to X{+}A$	real A vs random
Fold equal	$17.76 \to 15.56$; $\Delta = -2.20\ [-6.04, 1.40]$	$14.40 \to 12.74$; $\Delta = -1.66\ [-4.41, 0.86]$	$20/20$ both
Test-size weighted	$21.46 \to 15.60$; $\Delta = -5.86\ [-9.76, -0.09]$	$24.96 \to 20.62$; $\Delta = -4.34\ [-7.31, 0.40]$	$20/20$ both

primary target is Chl-a. After the pre-specified support screen, 583 samples remain: 126 from three HyperOCR campaigns and 457 from four RAMSES campaigns. The baseline is X =400–600 nm and the added measurement is A =602–800 nm. All X, $X + A$, and randomized-A conditions use identical samples and leave-one-campaign-out splits.

The held-out risk evidence is strongest under test-size weighting: ΔMAE is -5.86 with interval $[-9.76, -0.09]$. The estimated Mean K-Gap point estimate also decreases, but both campaign-bootstrap intervals for $\Delta\widehat{G}_{\text{mean}}$ cross zero. We therefore do not call the real-data gap movement certified. What is stable under the frozen protocol is the randomized comparison: the real extension is more favorable than every one of the 20 null repeats for both MAE and $\widehat{G}_{\text{mean}}$ under both aggregations (finite one-sided empirical rank $1/21$). This is a randomization diagnostic, not a causal test.

The heterogeneity is informative rather than hidden. On RAMSES, fold-equal ΔMAE is -4.94 with interval $[-9.12, -0.81]$ and the test-size-weighted change is -7.51 with interval $[-10.63, -0.53]$; HyperOCR is mixed under strong campaign-size imbalance. TSS improves MAE in six of seven folds but has a positive fold-equal mean because of one adverse RAMSES campaign. Leave-one-fold-out jackknife preserves the favorable aggregate Chl-a directions after every campaign deletion. Ridge does not reproduce the RF mean-gap direction and is retained as estimator sensitivity.

Before the spectral test, sample depth, water depth, month, and season were evaluated individually and jointly. None produced a stable simultaneous reduction in held-out MAE and the mean-gap proxy. These negative controls matter: not every available covariate is useful information for the task.

8.3 Does measurement extension move the tested risk curve?

On RAMSES Chl-a we train RF, histogram gradient boosting, MLP, and 1D-CNN at small, medium, and large capacity, using three fixed seeds and the same four leave-one-campaign-out folds. Adding A lowers mean MAE at every tested capacity for all four families. Medium $X + A$ also has lower mean MAE than Large X-only in every family, with differences -2.09 (1D-CNN), -2.79 (histogram boosting), -2.79 (MLP), and -4.04 (RF). The campaign/seed-level beat ratios are only 50–75%, so the result is a mean risk-curve shift over the tested range, not a claim that information universally dominates capacity.

9 Related work and scope

Identification, robust prediction, and Wasserstein geometry. Classical inverse theory asks whether latent quantities are uniquely and stably recoverable (Engl et al., 1996; Stuart, 2010); partial identification characterizes compatible sets when point identification fails (Manski, 2003). K-Gap is target-specific: it asks for a minimax prediction over environment-conditioned target laws rather than recovery of the entire latent world. Wasserstein DRO typically starts from a nominal distribution and a prescribed ambiguity neighborhood (Mohajerin Esfahani and Kuhn, 2018; Sinha et al., 2018); our set is instead induced by target mechanisms that remain compatible with the same observation. The abstract metric circumradius is not new. The new object is the conditional audit problem, together with the rectangular-envelope interpretation and the sample/coverage decomposition. A Wasserstein barycenter minimizes average distance (Agueh and Carlier, 2011); the K-Gap center minimizes worst-case distance and returns a common minimax prediction. The construction itself is metric-generic; Wasserstein is useful because the ground metric encodes target geometry, W_1 admits the Kantorovich–Rubinstein decision interpretation, and scalar centers are tractable through quantiles (Peyré and Cuturi, 2019; Villani, 2009; Chemseddine et al., 2025).

Distribution shift and measurement value. Domain adaptation, invariant learning, group robustness, and related methods optimize predictors under shift (Ben-David et al., 2010; Courty et al., 2017; Arjovsky et al., 2019; Sagawa et al., 2020; Gulrajani and Lopez-Paz, 2021; Koh et al., 2021); K-Gap is complementary because it audits the information set before choosing such a learner. The safe-refinement result is related to comparisons of statistical experiments (Blackwell, 1953). Hyperspectral regression provides a concrete stress test because water-quality targets are often inferred through indirect spectral correlations (Niu et al., 2021; He et al., 2025), with strong cross-site heterogeneity (Lechevallier et al., 2025) and diverse sensors and campaigns in GLORIA (Lehmann et al., 2023).

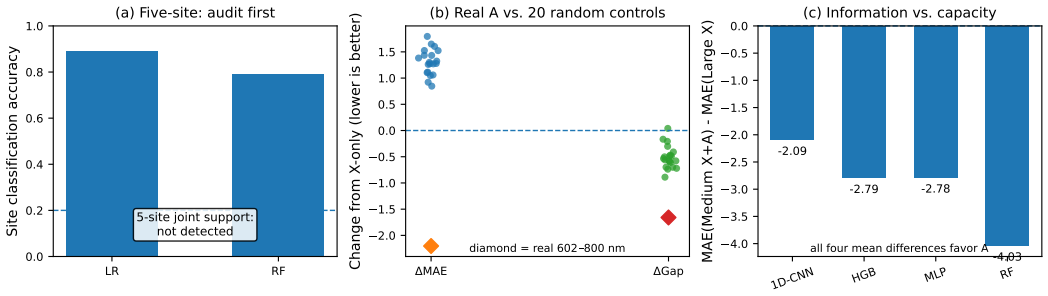

Figure 4: Real-data audit. **(a)** The frozen Five-site support diagnostic returns inconclusive rather than a gap estimate. **(b)** On GLORIA, the real 602–800 nm extension is more favorable than all 20 randomized-information controls for Chl-a/RF. **(c)** Across four model families, Medium $X + A$ has lower mean MAE than Large X-only, while campaign/seed-level outcomes remain heterogeneous.

Scope of the claims. Three boundaries are explicit. First, the exact robust-risk identity refers to the declared rectangular envelope; Proposition 3 shows how tighter globally coupled mechanism classes sit below it. Second, deployment beyond observed environments needs a scientific model of environment coverage and regularity; otherwise the audit is restricted to the declared observed family. Third, the GLORIA result is intentionally mean-level and empirical: its purpose is to test measurement value under matched samples, not to claim a certified distributional deployment K-Gap. These distinctions keep the theory and the real-data evidence on the same scale.

Four interventions, four mathematical objects. The framework is useful because the common responses to persistent error do not act on the same term. Increasing model capacity enlarges the predictor class $\mathcal{H}_k$ but leaves the declared compatible laws unchanged. Repeated samples can tighten η but do not cover unseen mechanisms. New environments can shrink h_m but do not, by themselves, change the population family. A new measurement changes the information set and may change the compatible-law family itself. This separation is the operational content of the audit: the same observed error can call for different resources depending on which term is unresolved.

Claim-to-evidence map. The same distinction is useful when reading the experiments: not every panel is intended to certify the same object.

Claim	Evidence	Status
Rectangular K-Gap equals robust Bayes risk	Theorem 2; exact controlled construction	exact under the declared envelope
Tighter global coupling can reduce robust risk	Proposition 3	exact comparison, no deployment coupling inferred from data
Sample error and environment coverage are distinct	Theorem 6; $4{,}500$ controlled coverage trials	certified in the controlled setting
Five-site common audit is supported	frozen overlap diagnostic	inconclusive under this protocol
GLORIA measurement improves prediction	matched LOCO risk, campaign bootstrap, 20 randomized controls	empirical held-out evidence
GLORIA population distributional K-Gap decreases	mean-level RF proxy only	not claimed

What would change the conclusion? The audit is designed so that negative outcomes remain informative. A failed overlap check prevents a gap claim; a wide calibration interval prevents certification; a real measurement that does not beat matched randomized information is not credited; and a capacity comparison is reported only over the tested range. Conversely, a tighter scientifically justified coupling class can replace the rectangular envelope and reduce the robust audit value without changing the finite-sample logic. This is why the framework is better read as a decision layer around prediction than as another predictor to benchmark against IRM, group DRO, or conformal methods.

10 Conclusion

Persistent prediction error can come from the learner or from the measurement, and these failures require different remedies. K-Gap turns compatible environment-conditioned target laws into a common minimax prediction and a declared-family ambiguity radius. The rectangular-envelope sandwich makes its conservatism explicit, while the finite-environment decomposition separates law-estimation error from mechanism-coverage error and yields a simpler observed-family certificate. Controlled tests recover the exact distributional radius and verify the finite-sample correction. On real data, the audit declines an unsupported Five-site comparison and shows that the GLORIA spectral extension lowers held-out Chl-a risk and beats all 20 randomized-information controls, while the mean-gap estimate is reported at its actual uncertainty level. Across four model families, the added measurement also shifts the tested mean risk curve beyond the medium-to-large capacity change, without implying uniform dominance. The central practical message is simple: before scaling a learner, determine which resource—capacity, repeated samples, environment coverage, or measurement information—addresses the bottleneck that is actually present.

REPRODUCIBILITY STATEMENT

All theoretical statements are accompanied by proofs in the appendix, with assumptions stated explicitly in the main text. For the controlled experiments, we specify the data-generating process, population K-Gap, model settings, sample sizes, and random seeds. For the real-data studies, we freeze the environment definitions, wavelength ranges, support rules, preprocessing, and outer leave-one-campaign-out splits before evaluating the reported comparisons. The supplementary material provides the experimental configurations, model-capacity definitions, fold-level predictions, randomized-A controls, bootstrap and jackknife outputs, and the scripts needed to reproduce the main tables and figures. All comparisons between X, $X + A$, and randomized-A use matched samples and identical outer splits. We report inconclusive support audits and estimator-sensitivity results rather than imputing unsupported quantities or selecting only favorable environments.

AI USE STATEMENT

In this paper/study, I only used AI tools to assist with figure creation, analysis of experimental conclusions, and writing polishing. Apart from these uses, AI was not used in any other part of the work, including but not limited to topic selection, research design, data collection, code writing, literature search and review, content generation, or translation. All AI-assisted content was reviewed, revised, and approved by me, and I take full responsibility for the views, data, and conclusions presented in this work.

REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Ruyan He, Zijun Lv, Yumiao Yang, and Sen Jia. Inversion of river water quality parameters from UAV hyperspectral data using a spatial–spectral attention CNN: A case study in Shenzhen. *Science of Remote Sensing*, 12:100282, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Pierre Lechevallier, W. Zhu, B. Shi, D. McCarthy, and J. Rieckermann. Non-contact hyperspectral monitoring of urban wastewater quality: Optimization of model calibration and performance. *Journal of Environmental Management*, 394:127217, 2025.

Moritz K. Lehmann et al. GLORIA - A globally representative hyperspectral in situ dataset for optical sensing of water quality. *Scientific Data*, 10:100, 2023.

Qin Li, Maria Oprea, Li Wang, and Yunan Yang. Inverse problems over probability measure space. *arXiv preprint arXiv:2504.18999*, 2025.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, NY, 2003.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

Chao Niu, Kun Tan, Xiuping Jia, and Xue Wang. Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery. *Environmental Pollution*, 286:117534, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

A PROOFS

A.1 PROOF OF THEOREM 2

For any measurable predictor q, let $\ell_q(x, \rho) = \mathsf{W}_p(q_x, \rho)$. The measurable-graph and compact-value assumptions imply measurability of $x \mapsto \sup_{\rho \in \mathcal{S}(x)} \ell_q(x, \rho)$. Every measurable selection is bounded by this pointwise supremum, while measurable ε-maximizing selections give the reverse inequality up to ε. Taking the infimum over measurable q, using measurable ε-centers, and letting $\varepsilon \downarrow 0$ yields eq. (5).

A.2 PROOF OF PROPOSITION 3

Because $\mathfrak{M} \subseteq \mathrm{Sel}(\mathcal{S})$, for every q,

$$\sup_{m \in \mathfrak{M}} \int \mathsf{W}_p(q_x, \rho_x^m) d\nu(x) \leq \sup_{\rho(\cdot) \in \mathrm{Sel}(\mathcal{S})} \int \mathsf{W}_p(q_x, \rho_x) d\nu(x).$$

Taking the infimum over q and applying Theorem 2 gives $R^\star(\mathfrak{M}) \leq \mathsf{G}$. If each fixed-environment mechanism m_e belongs to $\mathfrak{M}$, then $\sup_e \int \mathsf{W}_p(q_x, \rho_e(x)) d\nu(x) \leq \sup_{m \in \mathfrak{M}} \int \mathsf{W}_p(q_x, \rho_x^m) d\nu(x)$ for every q; taking infima gives the left inequality in eq. (6). The right equality for the rectangular envelope is Theorem 2.

A.3 PROOF OF COROLLARY 4

The rectangular robust objective is 1-Lipschitz in q under the integrated W_p metric. Density of $\cup_k \mathcal{H}_k$ implies that every measurable ε-center is approximated by some $q_k \in \mathcal{H}_k$ for sufficiently large k. Hence $R_k^\star \geq \mathsf{G}$ and $R_k^\star \leq \mathsf{G} + 2\varepsilon$ eventually, so $R_k^\star \downarrow \mathsf{G}$. For fixed k, empirical risk consistency gives $\widehat{R}_{k,n} \to R_k^\star$ as $n \to \infty$, and the stated iterated limit follows. A joint sequence requires uniform control over the growing classes.

A.4 PROOF OF THEOREM 6

Let $r_m(x) = \inf_q \max_j \mathsf{W}_p(q, \rho_{e_j}(x))$. Since the observed set is contained in the admissible family, $r_m(x) \leq \mathsf{G}(x)$. The triangle inequality gives $|\widehat{\mathsf{G}}_m - r_m| \leq \eta$, proving the lower bound. For any e, choose an observed e_j within h_m. For an ε-center $\widehat{q}$ of the estimated laws,

$$\mathsf{W}_p(\widehat{q}, \rho_e) \leq \mathsf{W}_p(\widehat{q}, \widehat{\rho}_j) + \mathsf{W}_p(\widehat{\rho}_j, \rho_{e_j}) + \mathsf{W}_p(\rho_{e_j}, \rho_e) \leq \widehat{\mathsf{G}}_m + \eta + L h_m + \varepsilon.$$

Taking the supremum over e, the infimum over q, and $\varepsilon \downarrow 0$ gives eq. (8).

A.5 PROOF OF COROLLARY 7

If the declared admissible family is exactly $\{e_1, \ldots, e_m\}$, every admissible environment is observed and $h_m(x) = 0$. Substituting into eq. (8) gives eq. (9) pointwise and after integration.

A.6 PROOF OF PROPOSITION 8

Under eq. (11), the target-law family after observing (x, a) is the subset $\{\rho_e(x) : e \in \mathcal{E}_{x,a}\}$ of the family available under x alone. Minimizing the maximum distance over a subset cannot increase the Chebyshev radius, so $\mathsf{G}_A(x, a) \leq \mathsf{G}(x)$ pointwise. Integration preserves the inequality.

A.7 PROOF OF PROPOSITION 9

Construct two meta-worlds that coincide on every observed environment and every observed sample distribution. In the first, all admissible deployment environments agree with the observed family. In the second, add an admissible deployment mechanism that is observationally indistinguishable on the training data but has a target law at positive Wasserstein distance. Any finite-sample algorithm has the same output distribution in both worlds, whereas the deployment K-Gaps differ. No estimator can therefore be uniformly consistent over both worlds.

Table 2: Fold-level Chl-a/RF measurement effects retained for transparency. Negative values favor the real 602–800 nm extension. The table shows why the aggregate result is reported with campaign-level uncertainty rather than as uniform improvement.

Sensor	Held-out campaign	n	ΔMAE	$\Delta\widehat{G}_{\mathrm{mean}}$
HyperOCR	Greb	10	-0.96	-1.23
HyperOCR	CST	9	$+5.53$	-0.92
HyperOCR	LK	107	-0.24	-0.21
RAMSES	Anstee	106	-3.06	-2.36
RAMSES	Barbosa	34	-6.71	$+4.20$
RAMSES	Higa	53	$+1.16$	-2.58
RAMSES	Jamet	264	-11.13	-8.50

B Additional real-data diagnostics

Five-site replication and abstention. The published PLS pipeline was reproduced before the support audit. Site classification from the shared 400–798 nm spectrum reached 0.89 accuracy with logistic regression and 0.79 with random forest on the balanced subset. Under the frozen fixed-cell rule, neither PCA features nor the raw 200-dimensional spectrum produced a cell with sufficient calibration support from all five sites. This is retained as an applicability audit rather than discarded as a failed benchmark.

Weak metadata interventions on GLORIA. Sample depth, water depth, month, and season were tested individually and jointly before the spectral-extension experiment. None produced a stable simultaneous reduction in held-out MAE and estimated Mean K-Gap across the frozen campaign folds. These negative controls motivate the criterion used in the main experiment: a mean-gap point estimate is reported together with held-out risk, campaign uncertainty, and randomized-information controls rather than being interpreted on its own.

Spectral-intervention robustness. For Chl-a+RF, leave-one-fold-out jackknife preserves $\Delta\mathrm{MAE} < 0$ and $\Delta\widehat{G}_{\mathrm{mean}} < 0$ for all seven deletions under fold-equal and test-size-weighted aggregation. Resampling the seven held-out campaigns gives 95% intervals $[-6.04, 1.40]$ and $[-9.76, -0.09]$ for fold-equal and test-size-weighted ΔMAE, and $[-4.41, 0.86]$ and $[-7.31, 0.40]$ for the corresponding $\Delta\widehat{G}_{\mathrm{mean}}$. The real extension is more favorable than all 20 frozen randomized-A repeats for both statistics under both aggregations, giving finite empirical rank $1/21$. RAMSES is favorable at the sensor level for MAE, HyperOCR is mixed under strong campaign-size imbalance, and TSS has aggregation-dependent MAE. Ridge is retained as estimator sensitivity because it does not reproduce the RF gap direction for Chl-a.

Scaling details. The RAMSES-Chl-a comparison uses four campaigns, three fixed seeds, four model families, three capacity levels, and matched $X/X + A$ splits. Across the 12 campaign/seed pairs, Small/Medium $X + A$ beats the corresponding Large X-only model in 50–75% of cases depending on family and capacity. These curves are descriptive evidence of risk-curve shift, not certified asymptotic convergence.

C Protocol details

Finite-sample certificate computation. The scalar certificate uses a split that keeps support construction separate from target-law calibration. A representation partition is fixed from training inputs without consulting target-test labels. Where calibration targets can be treated as conditionally i.i.d. within an environment–cell pair, their empirical CDF gives the DKW radius in eq. (10); the one-dimensional Chebyshev center is then computed on a common quantile grid. Cells failing the minimum calibration count remain inconclusive. This route is used for controlled certificate checks and is not asserted for correlated GLORIA campaigns.

GLORIA panel construction. The fixed-sensor intervention uses 583 support-screened samples. HyperOCR contributes 126 samples from three campaigns (9, 10, and 107), and RAMSES contributes 457 samples from four campaigns (34, 53, 106, and 264). Each sensor is evaluated with leave-one-campaign-out splits. The real intervention is always 602–800 nm added to the 400–600 nm baseline observation; the band boundary, campaigns, and target are not reselected after test results are seen. The 20 permuted-A controls preserve each added band's marginal distribution while breaking sample correspondence.

Robustness summaries. For the primary Chl-a/RF analysis we report fold-equal, test-size-weighted, and sensor-stratified summaries. Each held-out campaign uses 1,000 paired bootstrap resamples for MAE uncertainty, and a leave-one-fold-out jackknife checks whether a single campaign determines the aggregate direction. HyperOCR is reported as mixed rather than pooled away; TSS is retained as an environment-sensitive auxiliary target; Ridge is kept as estimator sensitivity.

Capacity comparison. The RAMSES-Chl-a study uses four model families (RF, histogram gradient boosting, MLP, and 1D-CNN), three capacity levels, three fixed seeds, and the same four outer campaign folds for X and $X + A$. The comparison is descriptive: it asks whether the tested increase in model capacity reproduces the risk shift caused by the additional measurement. It is not used to estimate the population K-Gap or to claim real-data asymptotic convergence.