SATRAP: ADAPTIVE RESIDUAL RETRIEVAL FOR AUGMENTING SATELLITE EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in satellite foundation models have produced a growing family of general-purpose embeddings for Earth-observation tasks. Once a downstream predictor is fitted, however, its training residuals no longer contribute to new predictions. We find that these residuals are strongly structured: geographically close observations with similar satellite embeddings tend to exhibit compatible prediction errors. Therefore, we introduce Satellite Residual-Augmented Prediction (SATRAP), a residual retrieval framework that augments embedding-based prediction using a geographically indexed representation–residual memory. SATRAP restricts retrieval by geographic proximity, measures compatibility in the embedding space, and adaptively aggregates the retrieved residuals. Under squared loss, the standard conditional-expectation identity gives the conditional mean residual given location and embedding as the optimal correction. We further derive an upper bound on the error of approximating this correction from finitely many retrieved residuals, in terms of geographic radius, embedding mismatch, and effective retrieval size. Experiments across satellite embeddings, downstream predictors, and global classification and regression tasks show that our approach achieves the best overall performance among the compared methods, improving mean classification accuracy by 4.51 percentage points and mean regression R^2 by 0.131 over the base predictors. These results establish structured residual retrieval as an effective complement to embedding-based prediction.

1 INTRODUCTION

Recent advances in satellite foundation models have made *Earth embeddings* increasingly important for Earth-observation analysis (Brown et al., 2025; Feng et al., 2026; Wang et al., 2025). An Earth embedding is a georeferenced feature vector that condenses the multispectral, multitemporal, and spatial satellite observations of the land, oceans, and atmosphere into a reusable representation. Such representations may be generated on demand by a pretrained encoder or distributed as precomputed global embedding products. A lightweight predictor fitted on these embeddings can support land-cover mapping, agricultural and ecosystem monitoring, and the estimation of terrain and vegetation properties, without retraining the underlying satellite model for every task (Szwarcman et al., 2026; Jakubik et al., 2025; Li et al., 2026). For these tasks, observations with comparable spectral and spatial characteristics may correspond to different target values under different climatic, ecological, or terrain conditions. These conditions vary across regions, making geographic location relevant to the relationship between an embedding and its prediction target. Earth embeddings may already encode such regional differences, but whether downstream predictors fully use them remains unclear.

Our analysis shows that geographic information, e.g., absolute location, regional identity, climate, and environmental variables, is recoverable from learned embeddings (Fig. 1, left). Yet the presence of this information does not establish how fully a downstream predictor uses it. Removing embedding components predictive of location affects downstream performance similarly to removing random components of the same rank (Fig. 1, right), suggesting that their geographic content does not confer a distinct downstream contribution (see Appendix D for details). These findings motivate examining whether geographic information remains predictive of the errors left by the downstream model.

We therefore examine whether the prediction errors retain structure associated with location and embedding. The errors observed while fitting this predictor are typically discarded once training is

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.

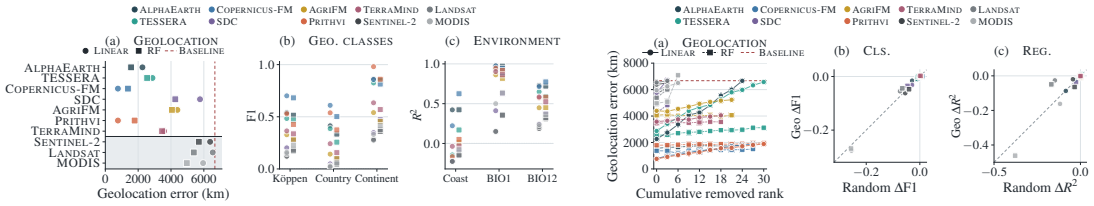


Figure 1: Geographic information in satellite representations (left) and downstream dependence (right). Geographic attributes are predictable from the representations. However, removing geographically informative components affects downstream performance similarly to random removal, suggesting that the identified geographic information is not fully used in downstream prediction.

complete. This procedure assumes that those errors contain no information that can be reused for future observations. However, we find that the residuals retain systematic structure not captured by the embedding-to-target mapping. In Fig. 2, residuals are most compatible when observations are geographically close and similar in the embedding space, and differences grow as either relation weakens. These patterns reveal a joint geographic–representation structure in prediction errors.

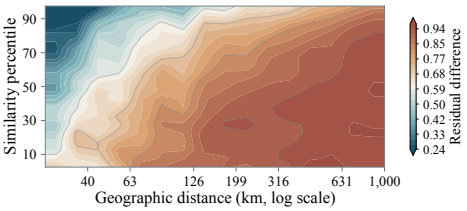


Figure 2: Residual compatibility by geographic distance and representation similarity.

To recover the predictive structure left unexplained by the fitted embedding-to-target mapping, we propose Satellite Residual-Augmented Prediction (SATRAP), which augments predictions from satellite embeddings through the adaptive retrieval of structured residuals. SATRAP associates training residuals with their locations and embeddings, forming an indexed residual memory attached to a fixed downstream predictor. For a new observation, geographic proximity identifies candidate residuals, embedding similarity determines their compatibility, and an adaptive scale controls their contribution. The resulting prediction is informed not only by the base estimate, but also by how the predictor has erred on comparable observations. For the finite residual memory used by SATRAP, we derive an error bound governed by geographic radius, embedding mismatch, and effective retrieval size. Experiments across learned Earth embeddings and satellite observations, global classification and regression tasks, multiple downstream predictor families, and alternative location representations show that SATRAP consistently improves prediction.

Our contributions are:

- We identify a gap between the geographic information encoded in satellite embeddings and its use in downstream prediction: embeddings encode this information through both linear and nonlinear structure, yet prediction errors retain geo-semantic structure with further predictive value.
- We propose SATRAP and its neurally enhanced variant SATRAP+ for retrieving structured residuals. We identify the optimal residual correction and derive a finite-retrieval error bound that explains the roles of geographic proximity, embedding similarity, and effective retrieval size.
- We evaluate across 10 satellite inputs, 6 global tasks, and 6 predictor families against 16 competing methods. Our approach achieves the best overall performance, with ablations and parameter analyses supporting the retrieval design.

2 Related Work

Satellite Representations. Earth embeddings are learned from global multisource data, including optical and radar imagery, terrain, atmospheric measurements, and environmental records. Earth representation learning has developed through both pretrained encoders and distributed embedding products. COPERNICUS-FM (Wang et al., 2025), GALILEO (Tseng et al., 2025), and TERRAMIND (Jakubik et al., 2025) learn transferable representations across diverse Earth-observation sources. Complementing these encoders, MAJOR TOM (Czerkawski et al., 2024) provides globally sampled embeddings, while ALPHAEARTH (Brown et al., 2025), TESSERA (Feng et al., 2026), and ESD (Chen et al., 2026) provide dense embedding layers. Beyond the information contained in individual embeddings, relationships among observations may offer further predictive value. This mo-

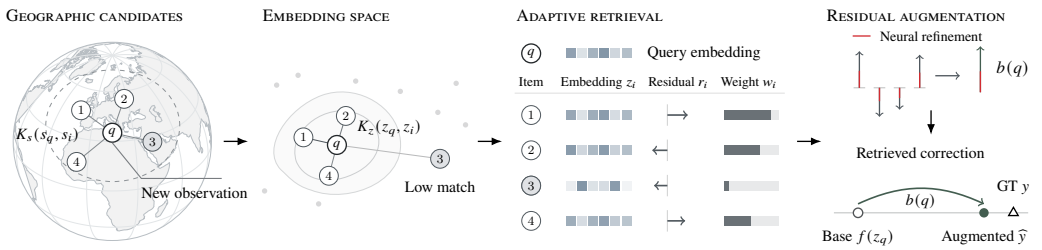


Figure 3: Structured residual retrieval for augmenting satellite embeddings.

tivates our investigation of how embeddings can be connected through geographic and representation structure to reduce downstream prediction errors.

Geographic Location Encoding. As discussed above, geographic relationships can also be represented through embeddings generated directly from coordinates. Location encoders map latitude and longitude into features describing spatial position and its associated regional, environmental, and social context. These representations can also be used to estimate errors left by a predictor. By encoding spatial relationships and associations with observed content, location representations provide informative features for modeling the prediction residuals of satellite embeddings. Coordinate-based encodings represent spatial relationships through direct coordinate transformations, multiscale periodic functions, and spherical representations. These approaches include Fourier features (Tancik et al., 2020), SPACE2VEC (Mai et al., 2020), SPHERE2VEC (Mai et al., 2023), and spherical-harmonic or implicit location representations (Ruß wurm et al., 2024; Cole et al., 2023). Pretrained encoders associate coordinates with geographically aligned observations to learn broader regional and environmental context. Their training sources span ground-level imagery in GEOCLIP (Vivanco Cepeda et al., 2023), satellite imagery in SATCLIP (Klemmer et al., 2025), RANGE (Dhakal et al., 2025), and TTE (Cher et al., 2026), and OpenStreetMap graphs in OSMGRAPHCLIP (Michail et al., 2026). Our approach achieves stronger overall performance than these methods. These results show that even coordinate-based location encodings specifically designed to capture spatial variation and location representations learned through large-scale global pretraining are less effective than SATRAP.

3 METHOD

Fig. 3 illustrates the central idea of SATRAP: recovering predictable residuals through joint geographic and semantic structure. We formulate the optimal correction, develop a geo-semantic retrieval estimator and its neurally enhanced variant, and derive an error bound that connects retrieval geometry with prediction improvement.

3.1 PROBLEM FORMULATION

Let $(x, s, y) \sim \mathbb{P}$ denote an observation, its geographic location, and its target drawn from the data distribution $\mathbb{P}$. A satellite encoder E produces $z = \mathrm{Pool}(E(x)) \in \mathbb{R}^D$, where D is the embedding dimension. We write $u(y) = y \in \mathbb{R}$, with $K = 1$, for regression and $u(y) = e_y \in \mathbb{R}^K$ for K-class classification, where e_y is the one-hot encoding of class y. A fitted predictor f maps z to the task output, and $q = (s, z)$ combines the location and embedding available for residual augmentation. An additive function $g(q)$ yields $P_K(f(z) + g(q))$, where P_K is the identity for regression and the Euclidean simplex projection for classification.

To treat both task types under the same objective, we define the population risk as the expected squared prediction loss

$$\mathcal{R}(g) = \mathbb{E}\Big[\|u(y) - P_K(f(z) + g(q))\|_2^2\Big]. \tag{1}$$

This is mean squared error for regression and multiclass Brier risk for classification. The corresponding unprojected residual is $r = u(y) - f(z)$, and the theoretical analysis uses $\widetilde{\mathcal{R}}(g) = \mathbb{E}\|r - g(q)\|_2^2$. Projection cannot increase the Brier loss; the task-specific expressions and projection argument are given in Appendices A and B.1.

Lemma 1 (Optimal residual augmentation) *Among all square-integrable functions of q, the population-optimal additive residual is the conditional mean residual. Moreover, the standard L^2 projection identity gives*

$$g^*(q) = \mathbb{E}[r \mid q], \qquad \widetilde{\mathcal{R}}(g) - \widetilde{\mathcal{R}}(g^*) = \mathbb{E}\|g(q) - g^*(q)\|_2^2. \tag{2}$$

Lemma 1 identifies the conditional mean residual, denoted by $m(q)$, as the target of augmentation. Its squared energy determines the maximum achievable risk reduction, while the mean squared approximation error determines the shortfall from this optimum. This relation connects the quality of residual estimation to downstream prediction and provides a common objective for the retrieval methods and location representations considered below.

Location encoders provide another representation for predicting the same residual target. Let e denote a location encoder and $e(s)$ the embedding that it assigns to coordinate s. Analytic encoders transform s through prescribed functions, whereas large-scale pretrained encoders such as SATCLIP (Klemmer et al., 2025) and TTE (Cher et al., 2026) can learn $e(s)$ from satellite imagery, ground-level imagery, maps, environmental observations, etc. They can therefore be viewed as a pretrained form of the query $q = (s, z)$, with the learned relation between s and z compressed into $e(s)$. We therefore include these methods in the comparison with SATRAP.

3.2 SATELLITE RESIDUAL-AUGMENTED PREDICTION

Given a fitted predictor, SATRAP retains $\mathcal{M} = \{(s_i, z_i, r_i)\}_{i=1}^n$ as a residual memory of n observations, with $q_i = (s_i, z_i)$ and $r_i = u(y_i) - f(z_i)$. Each tuple is an empirical observation of the unknown field $m(q)$, indexed jointly by geographic location and satellite representation. Geographic separation is measured by great-circle distance d_s, while d_z is angular distance between training-normalized embeddings; complete definitions are given in Appendix A.2.

For a query $q = (s_q, z_q)$, SatRAP first restricts estimation to $\mathcal{C}_M(s_q)$, the index set of the M geographically closest memory entries. Unless otherwise specified, retrieval sums run over this candidate set. This geographic support prevents representation similarity alone from retrieving arbitrarily distant residuals.

Definition 1 (SatRAP residual retrieval) Let $K_s(s_q, s_i) \geq 0$ be a location kernel and $K_z(z_q, z_i) \geq 0$ an embedding kernel. SatRAP assigns candidate i a normalized weight satisfying

$$w_i(q) \propto K_s(s_q, s_i) K_z(z_q, z_i), \qquad \sum_i w_i(q) = 1, \tag{3}$$

and estimates the conditional residual by

$$b(q) = \sum_{i \in \mathcal{C}_M(s_q)} w_i(q) r_i. \tag{4}$$

Thus, geography determines where residual evidence may be borrowed, while the embedding determines how that evidence is distributed within the geographic support. The result $b(q)$ is a locally normalized, representation- conditioned estimator of the optimal residual in Lemma 1. Its finite-memory estimation error $\|b(q) - m(q)\|_2^2$ is analyzed in Section 3.3 in terms of geographic radius, representation mismatch, and effective retrieval size.

SatRAP adapts the correction magnitude to the retrieval context of each query through the shrinkage function $\eta_\theta(q) = \mathrm{sigmoid}(\theta_0 + \boldsymbol{\theta}^\top c(q))$, where $\theta = (\theta_0, \boldsymbol{\theta})$ comprises an intercept θ_0 and a coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^5$. The standardized vector $c(q) \in \mathbb{R}^5$ summarizes retrieval range, local support, effective retrieval size, and residual dispersion. Its parameters are fitted directly against the residual objective:

$$\theta^* = \arg\min_\theta \frac{1}{|\mathcal{I}|K} \sum_{j \in \mathcal{I}} \|r_j - \eta_\theta(q_j) b(q_j)\|_2^2 + \frac{\lambda_\eta}{5} \|\boldsymbol{\theta}\|_2^2, \tag{5}$$

where $\mathcal{I}$ indexes the fitting observations and $\lambda_\eta \geq 0$ regularizes the coefficient vector. Each observation is excluded from its own retrieval set. Full feature definitions and estimation details

are provided in Appendix A.3. The fitted shrinkage function scales the retrieved residual before it augments the base prediction.

The construction above does not require a particular pair of kernels. In this work, the location kernel decays exponentially with normalized geographic distance, while the embedding kernel increases exponentially with $\mathrm{sim}(z_q, z_i)$, the cosine similarity computed after fitting-data standardization:

$$K_s(s_q, s_i) = \exp\left[-\frac{d_s(s_q, s_i)}{h(q)}\right], \qquad K_z(z_q, z_i) = \exp[\beta\,\mathrm{sim}(z_q, z_i)]\,. \tag{6}$$

Here $h(q) = \max\{d_{k_a}(q), h_{\min}\}$ adapts the spatial scale to local sampling density. The distance $d_{k_a}(q)$ is measured to the k_a-th geographically nearest candidate, where $1 \le k_a \le M$ sets the bandwidth rank rather than the number of contributing entries; $h_{\min} > 0$ is the minimum bandwidth. The coefficient $\beta \ge 0$ controls representation selectivity. The equivalent log-score is $a_i(q) = -d_s(s_q, s_i)/h(q) + \beta\,\mathrm{sim}(z_q, z_i)$.

Parameterized residual retrieval. The two analytic kernels encode complementary structure: K_s restricts retrieval according to geographic proximity, whereas K_z measures agreement between satellite representations. The most direct route to end-to-end training is to replace K_s and K_z entirely with neural networks and learn the retrieval rule from the downstream loss. However, we find that this replacement discards the structure encoded by SATRAP: the explicit geographic decay and representation-agreement term are removed, and the resulting model performs substantially worse. This motivates SATRAP+, which preserves the analytic score $a_i(q)$ as a structured retrieval prior and learns task-adaptive neural affinities over the joint geographic–representation space directly from the downstream loss. The comparison in Table 3 shows that retaining the SATRAP retrieval structure yields stronger performance than parameterizing the entire retrieval rule.

Two compact two-layer GELU MLPs, $\phi_s(\cdot;\psi_s)$ and $\phi_z(\cdot;\psi_z)$, with learnable network parameters ψ_s and ψ_z, respectively, map the unit-sphere coordinates of s and the standardized satellite embedding z to normalized location and representation keys. With learned scalar coefficients α_s and α_z, the parameterized retrieval score is

$$a_i^+(q;\psi) = a_i(q) + \alpha_s \phi_s(s_q;\psi_s)^\top \phi_s(s_i;\psi_s) + \alpha_z \phi_z(z_q;\psi_z)^\top \phi_z(z_i;\psi_z), \tag{7}$$

where $\psi = (\psi_s, \psi_z, \alpha_s, \alpha_z)$. The candidate set remains $\mathcal{C}_M(s_q)$, but its residuals are reweighted according to

$$w_i^+(q;\psi) = \frac{\exp a_i^+(q;\psi)}{\sum_{\ell\in\mathcal{C}_M(s_q)} \exp a_\ell^+(q;\psi)}, \qquad b_\psi^+(q) = \sum_{i\in\mathcal{C}_M(s_q)} w_i^+(q;\psi) r_i. \tag{8}$$

The kernel parameters are learned directly from the downstream objective. For the fitting queries $\mathcal{I}$, with self-retrieval excluded, we minimize

$$\psi^* = \arg\min_\psi \frac{1}{|\mathcal{I}|} \sum_{j\in\mathcal{I}} \left\| u(y_j) - P_K\Big(f(z_j) + b_\psi^+(q_j)\Big)\right\|_2^2. \tag{9}$$

The coefficients α_s and α_z are initialized at zero, so the initial retrieval is exactly SATRAP. With $w_i^+ \ge 0$ and $\sum_i w_i^+ = 1$, $b_{\psi^*}^+(q)$ remains a convex combination of the stored residuals, with their contributions adapted to the downstream task.

3.3 Residual Retrieval Error Bound

To characterize the finite-memory approximation, write the base residual as $r = m(q) + \varepsilon$, where $m(q) = \mathbb{E}[r \mid q]$ and ε is the remaining variation; for memory entry i, $\varepsilon_i = r_i - m(q_i)$. Assume that m is jointly Lipschitz continuous with respect to geographic and embedding distances, with constants L_s and L_z, respectively. The memory noise is assumed to be conditionally zero mean and uncorrelated across entries, with energy bounded by σ^2.

Define the weighted geographic radius $R_s(q) = \sum_i w_i(q) d_s(s_i, s)$, the weighted representation mismatch $R_z(q) = \sum_i w_i(q) d_z(z_i, z)$, and the effective retrieval size $N_{\mathrm{eff}}(q) = [\sum_i w_i(q)^2]^{-1}$.

Proposition 1 (Residual retrieval error) *Under the assumptions above, the following bound characterizes the accuracy with which a finite residual memory approximates the population-optimal correction:*

$$\mathbb{E}[\|b(q) - m(q)\|_2^2 \mid q, \{q_i\}] \leq [L_s R_s(q) + L_z R_z(q)]^2 + \frac{\sigma^2}{N_{\text{eff}}(q)}. \tag{10}$$

The bound follows from the decomposition

$$b(q) - m(q) = \sum_i w_i [m(q_i) - m(q)] + \sum_i w_i \varepsilon_i. \tag{11}$$

The two sums in Eq. (11) are the retrieval bias and retrieval noise, respectively. Their corresponding terms in Eq. (10) increase with geographic or representation mismatch and when retrieval weight is concentrated on few entries. The bound therefore exposes the trade-off encoded by SatRAP: restricting geographic support controls R_s; emphasizing representation agreement can reduce R_z, but may concentrate the weights and reduce N_{eff}. The adaptive bandwidth and representation coefficient balance retrieval bias and variance for each query. The bound provides a sufficient condition for risk reduction: the retrieval error bound must be smaller than the squared magnitude of the conditional mean residual. The complete proof and risk-gain identity are in Appendix B.2. Together, Lemma 1 and Proposition 1 connect the optimal residual correction to its finite-memory estimate in SATRAP.

4 RESULTS

Experiments cover 10 satellite inputs, 6 global tasks, 6 predictor families, and 5 mutually exclusive IID folds. All residual methods share the same base predictions and training residuals, and test labels are used only for evaluation. Residual baselines comprise XYZ, DIRECT, and SH-L15; WRAP (Mac Aodha et al., 2019); SPACE2VEC-G/T (Mai et al., 2020); SPHEREC/C+ and SPHEREM/M+ (Mai et al., 2023); and pretrained GEOCLIP (Vivanco Cepeda et al., 2023), SATCLIP (Klemmer et al., 2025), RANGE/RANGE+ (Dhakal et al., 2025), TTE (Cher et al., 2026), and OSMGC (Michail et al., 2026). Relative risk reduction is $\Delta_{\text{risk}}(g) = [\mathcal{R}(0) - \mathcal{R}(g)]/\mathcal{R}(0)$. To evaluate the observable implications of Proposition 1, we relate the geographic radius $R_s(q)$, representation mismatch $R_z(q)$, and effective retrieval size $N_{\text{eff}}(q)$ to the observed reduction in squared error before scaling and projection. Datasets, predictors, comparison methods, and the complete evaluation protocol are given in Appendix C.

4.1 MAIN COMPARISON AND ABLATION STUDIES

We evaluate whether the joint residual structure observed in Fig. 2 translates into predictive gains and test the contribution of each retrieval component.

Table 1 and Fig. 4 show that SATRAP improves predictions from both learned embeddings and raw observations, with SATRAP+ further improving OA, R^2, and risk reduction. In Fig. 4, pretrained methods generally attain larger average gains than non-pretrained methods, while SATRAP and SATRAP+ rank highest. SATRAP+ reduces mean risk by 14.68% with embeddings and 26.38% with raw observations, compared with the pretrained-method averages of 9.81% and 19.36%.

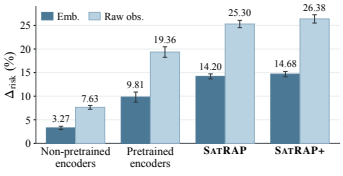


Figure 4: Mean test Δ_{risk}.

Table 1 details these improvements for individual tasks. On elevation, SH-L15, SATCLIP (Klemmer et al., 2025), RANGE+ (Dhakal et al., 2025), and TTE (Cher et al., 2026) already yield substantial gains, yet SATRAP+ reaches the highest R^2: 0.8552 with embeddings and 0.7690 with raw observations, compared with TTE's 0.8444 and 0.7618. On canopy height and biomass, SATRAP+ reduces risk by 3.2% and 9.4% with embeddings, exceeding the strongest competing reductions of 0.4% from WRAP (Mac Aodha et al., 2019) and 6.4% from RANGE+, respectively. With raw observations, these reductions reach 13.6% and 24.5%. SATRAP+ attains the highest OA and risk reduction across all three classification tasks. On FROM-GLC10, RANGE+ slightly exceeds SATRAP in OA.

Table 1: Comparison of methods: test OA and R^2; $(\cdot)$ reports Δ_{risk}. **Bold** and underline mark the best and second-best values, respectively. † denotes a pretrained location encoder.

Input	Method	OA (Δ_{risk})			R^2 (Δ_{risk})		
		ESRI	WorldCover	FROM-GLC10	Canopy	AGB	DEM
Emb.	EMB. ONLY	0.6926	0.5909	0.6274	0.4890	0.5480	0.6589
	XYZ	0.6976 (+1.6%)	0.5933 (+0.7%)	0.6317 (+0.9%)	0.4896 (+0.1%)	0.5490 (+0.2%)	0.6607 (+0.5%)
	DIRECT	0.6973 (+1.6%)	0.5931 (+0.6%)	0.6301 (+0.6%)	0.4895 (+0.1%)	0.5488 (+0.2%)	0.6604 (+0.4%)
	SH-L15	0.7074 (+3.9%)	0.6081 (+3.5%)	0.6487 (+5.0%)	0.4844 (−1.0%)	0.5545 (+1.5%)	0.8019 (+42.1%)
	WRAP (Mac Aodha et al., 2019)	0.6976 (+1.7%)	0.5935 (+0.8%)	0.6320 (+1.0%)	0.4912 (+0.4%)	0.5497 (+0.4%)	0.6628 (+1.2%)
	SPACE2VEC-G (Mai et al., 2020)	0.7023 (+2.7%)	0.6002 (+2.0%)	0.6396 (+2.7%)	0.4893 (+0.1%)	0.5502 (+0.8%)	0.7108 (+15.3%)
	SPACE2VEC-T (Mai et al., 2020)	0.7048 (+3.3%)	0.6027 (+2.4%)	0.6440 (+3.6%)	0.4901 (+0.2%)	0.5545 (+1.5%)	0.7343 (+22.2%)
	SPHEREC (Mai et al., 2023)	0.7020 (+2.5%)	0.5994 (+1.7%)	0.6394 (+2.7%)	0.4872 (−0.3%)	0.5534 (+1.4%)	0.6997 (+12.2%)
	SPHERECC+ (Mai et al., 2023)	0.7037 (+2.9%)	0.6037 (+2.6%)	0.6440 (+3.7%)	0.4856 (−0.7%)	0.5493 (+0.6%)	0.7313 (+21.5%)
	SPHEREM (Mai et al., 2023)	0.6982 (+1.8%)	0.5934 (+0.7%)	0.6309 (+0.9%)	0.4873 (−0.3%)	0.5455 (−0.5%)	0.6677 (+2.6%)
	SPHEREM+ (Mai et al., 2023)	0.7001 (+2.3%)	0.5968 (+1.4%)	0.6359 (+2.0%)	0.4908 (+0.3%)	0.5514 (+0.9%)	0.6934 (+10.2%)
	GEOCLIP† (Vivanco Cepeda et al., 2023)	0.7032 (+1.8%)	0.6043 (+1.7%)	0.6416 (+2.0%)	0.4493 (−7.7%)	0.4830 (−13.8%)	0.7922 (+39.3%)
	SATCLIP† (Klemmer et al., 2025)	0.7090 (+4.3%)	0.6109 (+4.0%)	0.6512 (+5.4%)	0.4855 (−0.7%)	0.5618 (+3.3%)	0.8153 (+46.0%)
	RANGE† (Dhakal et al., 2025)	0.7115 (+5.2%)	0.6137 (+4.7%)	0.6522 (+5.7%)	0.4875 (−0.2%)	0.5666 (+4.5%)	0.8328 (+51.0%)
	RANGE+† (Dhakal et al., 2025)	0.7120 (+5.5%)	0.6150 (+5.0%)	0.6547 (+6.3%)	0.4902 (+0.3%)	0.5752 (+6.4%)	0.8379 (+52.5%)
	TTE† (Cher et al., 2026)	0.7096 (+4.1%)	0.6122 (+4.0%)	0.6535 (+5.4%)	0.4723 (−3.2%)	0.5372 (−1.9%)	0.8444 (+54.5%)
	OSMGC† (Michail et al., 2026)	0.7086 (+4.2%)	0.6101 (+4.0%)	0.6495 (+5.2%)	0.4872 (−0.4%)	0.5684 (+4.8%)	0.8023 (+42.1%)
	SATRAP (Ours)	0.7127 (+5.7%)	0.6154 (+5.1%)	0.6539 (+6.3%)	0.5051 (+3.1%)	0.5878 (+8.9%)	0.8499 (+56.0%)
	SATRAP+ (Ours)	**0.7132** (+6.1%)	**0.6159** (+5.3%)	**0.6549** (+6.6%)	**0.5053** (+3.2%)	**0.5897** (+9.4%)	**0.8552** (+57.6%)
Raw obs.	OBS. ONLY	0.6007	0.4883	0.5485	0.4055	0.4508	0.2176
	XYZ	0.6210 (+4.2%)	0.5005 (+1.9%)	0.5673 (+2.9%)	0.4102 (+0.8%)	0.4563 (+1.0%)	0.2229 (+0.7%)
	DIRECT	0.6209 (+4.1%)	0.4998 (+1.8%)	0.5617 (+2.1%)	0.4108 (+0.9%)	0.4577 (+1.2%)	0.2243 (+0.8%)
	SH-L15	0.6626 (+11.7%)	0.5585 (+9.4%)	0.6261 (+12.2%)	0.4419 (+6.0%)	0.5269 (+14.1%)	0.6397 (+54.1%)
	WRAP (Mac Aodha et al., 2019)	0.6222 (+4.5%)	0.5028 (+2.3%)	0.5713 (+3.5%)	0.4205 (+2.5%)	0.4669 (+2.9%)	0.2360 (+2.4%)
	SPACE2VEC-G (Mai et al., 2020)	0.6439 (+8.1%)	0.5279 (+5.5%)	0.5984 (+7.3%)	0.4273 (+3.6%)	0.4906 (+7.8%)	0.3921 (+22.4%)
	SPACE2VEC-T (Mai et al., 2020)	0.6495 (+9.2%)	0.5382 (+6.6%)	0.6109 (+8.9%)	0.4295 (+4.0%)	0.5056 (+10.0%)	0.4494 (+29.7%)
	SPHEREC (Mai et al., 2023)	0.6423 (+7.6%)	0.5298 (+5.4%)	0.6028 (+7.9%)	0.4245 (+3.2%)	0.5041 (+9.8%)	0.3645 (+19.0%)
	SPHERECC+ (Mai et al., 2023)	0.6515 (+9.3%)	0.5432 (+7.1%)	0.6148 (+9.6%)	0.4281 (+3.7%)	0.5107 (+11.3%)	0.4582 (+31.0%)
	SPHEREM (Mai et al., 2023)	0.6252 (+4.9%)	0.5048 (+2.3%)	0.5740 (+3.7%)	0.4171 (+1.9%)	0.4663 (+2.8%)	0.2626 (+5.7%)
	SPHEREM+ (Mai et al., 2023)	0.6326 (+6.2%)	0.5138 (+4.0%)	0.5898 (+6.0%)	0.4249 (+3.2%)	0.4823 (+6.0%)	0.3429 (+16.1%)
	GEOCLIP† (Vivanco Cepeda et al., 2023)	0.6553 (+9.3%)	0.5469 (+6.6%)	0.6100 (+7.6%)	0.3928 (−2.2%)	0.4494 (+0.3%)	0.6250 (+52.4%)
	SATCLIP† (Klemmer et al., 2025)	0.6696 (+13.0%)	0.5662 (+10.7%)	0.6348 (+13.4%)	0.4436 (+6.3%)	0.5343 (+15.6%)	0.6817 (+59.4%)
	RANGE† (Dhakal et al., 2025)	0.6783 (+15.3%)	0.5737 (+12.2%)	0.6347 (+13.7%)	0.4468 (+7.0%)	0.5434 (+17.4%)	0.7250 (+64.9%)
	RANGE+† (Dhakal et al., 2025)	0.6786 (+15.5%)	0.5761 (+12.7%)	0.6408 (+14.6%)	0.4506 (+7.6%)	0.5588 (+20.2%)	0.7364 (+66.4%)
	TTE† (Cher et al., 2026)	0.6786 (+14.3%)	0.5735 (+11.9%)	0.6401 (+14.0%)	0.4356 (+5.1%)	0.5141 (+12.2%)	0.7618 (+69.6%)
	OSMGC† (Michail et al., 2026)	0.6674 (+12.4%)	0.5634 (+10.2%)	0.6289 (+12.5%)	0.4416 (+6.0%)	0.5399 (+16.7%)	0.6429 (+54.4%)
	SATRAP (Ours)	0.6871 (+17.4%)	0.5861 (+14.5%)	0.6492 (+16.6%)	0.4773 (+12.0%)	0.5751 (+23.0%)	0.7523 (+68.4%)
	SATRAP+ (Ours)	**0.6906** (+18.2%)	**0.5877** (+14.8%)	**0.6501** (+16.8%)	**0.4868** (+13.6%)	**0.5831** (+24.5%)	**0.7690** (+70.5%)

Taking elevation as an example, predictors can leave systematic errors where pronounced terrain variation is not fully resolved. Such errors may recur across mountain ranges and highlands rather than occur in isolation. In Fig. 5, averaged over 10 satellite inputs and 6 predictors, SATRAP shows pronounced improvements over XYZ and SPHERE2VEC (Mai et al., 2023) around the Tibetan Plateau and the Andes. Pretrained methods, including SATCLIP (Klemmer et al., 2025), RANGE+ (Dhakal et al., 2025), and TTE (Cher et al., 2026), benefit from geographic representations learned through large-scale pretraining and already yield substantial improvements in elevation prediction; SATRAP further reduces errors across parts of western North America and the East African highlands. These results demonstrate the superiority of our method.

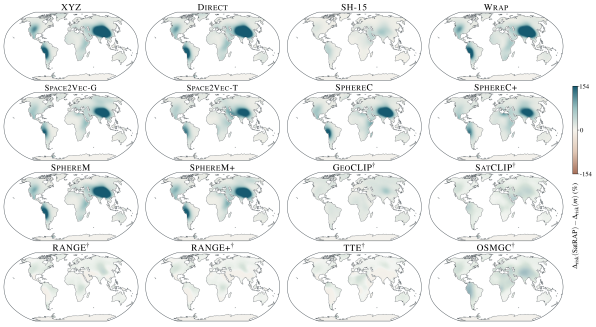


Figure 5: SATRAP elevation Δ_{risk} improvement.

Table 2 shows why geographic proximity and embedding similarity need to be considered together. Geography-only retrieval outperforms embedding-only retrieval, indicating that geographic proximity is particularly informative for identifying relevant residuals. Adding embedding similarity raises classification risk reduction from 6.93% to 8.63% and regression risk reduction from 18.82% to 22.35%. Shuffling the similarity scores reduces performance below geography-only retrieval, showing that the improvement depends on correctly matching representations to the retrieved residuals, rather than merely introducing additional weights. Geographic proximity therefore identifies useful observations, while embedding similarity helps identify relevant residuals.

Table 3 further examines whether this retrieval structure can be learned entirely from the downstream objective. FULL-PARAM. replaces the analytic kernels with neural similarities, removing the explicit geographic decay and original embedding-similarity terms. Regression risk reduction falls from 22.35% to 10.62%, with classification performance also declining. SATRAP+ retains these terms

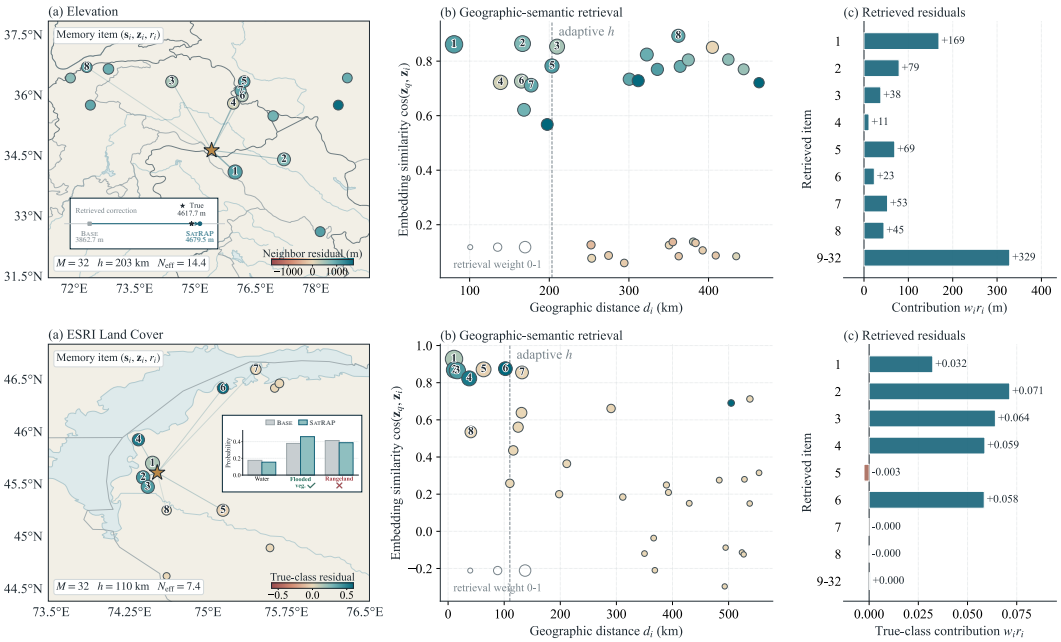


Figure 6: SATRAP retrieval for elevation (top) and land cover (bottom).

and improves upon SATRAP. This comparison supports preserving the geographic and representation structure during parameterization rather than learning the entire retrieval rule from scratch.

Table 2: Geographic and representation retrieval.

Method	OA (Δ_{risk})	R^2 (Δ_{risk})
GEO. ONLY	0.6477 (6.93%)	0.6159 (18.82%)
EMB. ONLY	0.6216 (2.90%)	0.5221 (4.68%)
SHUFFLED SIM.	0.6406 (5.60%)	0.6073 (17.24%)
SATRAP	**0.6547** (**8.63%**)	**0.6338** (**22.35%**)

Table 3: Analytic and neural retrieval.

Method	OA (Δ_{risk})	R^2 (Δ_{risk})
FULL-PARAM.	0.6357 (5.86%)	0.5586 (10.62%)
SATRAP	0.6547 (8.63%)	0.6338 (22.35%)
SATRAP+	**0.6558** (**8.97%**)	**0.6390** (**23.42%**)

4.2 RESIDUAL RETRIEVAL EXAMPLES

Fig. 6 illustrates how retrieval relevance depends jointly on geographic proximity and embedding similarity. Nearby observations can receive little weight when their representations differ, while more distant observations can remain influential. In the elevation example, item 8 remains among the leading contributions beyond the 203-km adaptive scale, which controls distance decay rather than imposing a hard cutoff. Panel (b) compares all 32 candidates, while panel (c) separates the contributions of the 8 highest-weight entries from those of the remaining candidates. Together, they show how retrieval relevance and residual magnitude determine the correction.

The resulting corrections improve both numerical estimates and class probabilities. For elevation, the estimate increases from 3862.7 to 4679.5 m, approaching the target value of 4617.7 m. For ESRI land cover, the probability of the true flooded-vegetation class rises from 37.78% to 45.91%, while rangeland decreases from 41.26% to 38.68%, correcting the predicted class. These examples illustrate how contributions of different magnitudes and signs combine to reduce prediction errors, without requiring every retrieved observation to support the same correction.

4.3 RETRIEVAL WITH LOCATION EMBEDDINGS

The geographic component of SATRAP need not be restricted to great-circle distance d_s. A location embedding $e(s)$ can instead induce the metric used to form the candidate set $\mathcal{C}_M(s_q)$ and compute the geographic kernel $K_s(s_q, s_i)$. Interestingly, pretrained location embeddings perform markedly better overall than their non-pretrained counterparts in Table 4. Most further improve SATRAP over

its geodesic formulation, whereas all non-pretrained alternatives reduce performance on both tasks. This suggests that pretrained regional and environmental context helps identify observations with compatible residuals. TTE (Cher et al., 2026) yields the largest classification gain, while RANGE+ (Dhakal et al., 2025) gives the largest regression gain; SATCLIP (Klemmer et al., 2025) also improves both tasks. These gains show that SATRAP effectively incorporates geographic structure learned through large-scale pretraining to further reduce prediction errors.

Table 4: Location-embedding retrieval; green denotes gains over geodesic.

Method	ESRI ΔOA (%)	AGB ΔR^2
SPACE2VEC-T	-0.05%	-0.0037
SPACE2VEC-G	-0.06%	-0.0063
SPHEREC	-0.04%	-0.0062
SPHERECC+	-0.05%	-0.0059
SPHEREM	-2.20%	-0.0513
SPHEREM+	-1.57%	-0.0396
GEOCLIP†	$+0.00\%$	-0.0085
SATCLIP†	$+0.10\%$	$+0.0032$
RANGE†	$+0.20\%$	-0.0020
RANGE+†	$+0.35\%$	$\mathbf{+0.0075}$
TTE†	$\mathbf{+0.36\%}$	$+0.0045$
OSMGC†	$+0.10\%$	$+0.0001$

4.4 RETRIEVAL GEOMETRY AND PERFORMANCE

Fig. 7 connects the parameters of SATRAP to prediction performance and the retrieval quantities in Proposition 1. The top row reports mean test Δ_{risk}; the bottom row shows the corresponding geographic radius R_s, with color indicating representation mismatch R_z and circle area indicating effective retrieval size N_{eff}. Increasing the candidate-pool size M admits more distant memory entries, whereas increasing the bandwidth rank k_a enlarges $h(q)$ and weakens geographic distance decay. Both changes increase R_s and N_{eff} in the bottom row. Yet test improvement peaks near $M = 32$ and $k_a = 6$–8, rather than at the largest effective retrieval size. This agrees with the competing terms in Eq. (10): averaging over more entries reduces the noise term σ^2/N_{eff}, but broader geographic support and greater representation mismatch can increase the bias term. Loss of locality can offset the benefit of averaging more residuals.

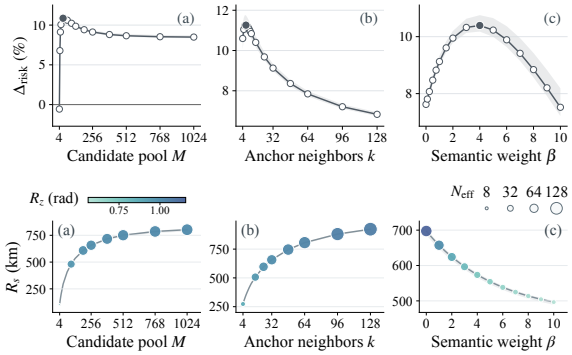


Figure 7: Sensitivity (top); retrieval geometry (bottom).

The semantic coefficient β changes this balance from the other direction. Increasing it strengthens the embedding kernel, lowering R_z while concentrating weight on fewer entries, as shown by the changing colors and smaller circles. Test improvement peaks near $\beta = 4$–5 and then declines despite closer representation matching. Thus, reducing mismatch alone does not ensure better prediction: a smaller N_{eff} increases the variance term in the bound. Together, the two rows support SATRAP's combination of local geographic support and selective representation weighting.

5 DISCUSSION AND CONCLUSION

This work shows that relationships among Earth embeddings retain predictive value beyond the fitted embedding-to-target mapping. Prediction residuals exhibit geo-semantic structure, which SATRAP recovers through adaptive retrieval. Our analysis derives a residual retrieval error bound involving geographic proximity, representation agreement, and effective retrieval size. SATRAP outperforms the compared methods overall, with SATRAP+ achieving further gains. These results support structured residual retrieval as a means of improving prediction from satellite embeddings.

The experiments also clarify which aspects of retrieval matter. Geographic proximity and embedding similarity are more effective together, and replacing their kernels entirely with learned similarities reduces performance. SATRAP+ instead learns a refinement of these kernels, suggesting that their structure remains useful even when retrieval is optimized for the task. The parameter results are consistent with the theoretical trade-off between retrieval bias and variance. Future work could investigate how structured residuals across diverse downstream tasks can guide the learning of Earth embeddings, incorporating task-specific predictive evidence into general-purpose representations.

Reproducibility Statement

The datasets and code for the comparison methods are publicly available for download. Random seeds are fixed for data splitting and model fitting to support reproducibility. The method definitions and implementation details are given in Section 3 and Appendix A. The assumptions and complete proofs are provided in Appendix B. Data sources, target preparation, data splits, and evaluation procedures are documented in Appendix C; predictor settings and location-encoder configurations are listed in Appendices C.4 and C.5. Code and derived evaluation summaries will be released.

Ethics Statement

This study uses existing satellite and environmental data products and does not involve experiments on human participants or animals.

AI Use Statement

In this work, we used generative AI tools to polish the manuscript, summarize existing literature, and assist with code implementation. We have not used generative AI tools to develop theoretical models, formulate mathematical claims, write proofs, propose hypotheses, design experiments, prepare data, generate synthetic datasets, or interpret results. Survey, interview, and human-subject data analysis tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Copernicus dem, 2022. URL http://dx.doi.org/10.5270/ESA-c5d3d65.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

S. Chen, J. Wang, S. Yuan, J. Li, Y. Xia, Y. Liao, J. Wei, J. Yuan, X. Xu, X. Zhu, P. Zhu, H. Zhang, Y. Zhou, H. Fu, H. Huang, B. Chen, F. Dai, and P. Gong. Democratizing planetary-scale analysis: an ultra-lightweight earth embedding database for accurate and flexible global land monitoring. *Earth System Science Data*, 18(7):5375–5398, 2026. doi: 10.5194/essd-18-5375-2026. URL https://essd.copernicus.org/articles/18/5375/2026/.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Daniel Cher, Hamza Iqbal, Eric Xing, Brian Wei, and Nathan Jacobs. Tessellating the earth. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 443–459, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37092-1.

Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6320–6342. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cole23a.html.

Mikolaj Czerkawski, Marcin Kluczek, and Jędrzej S. Bojanowski. Global and dense embeddings of earth: Major tom floating in the latent space, 2024. URL https://arxiv.org/abs/2412.05600.

Aayush Dhakal, Srikumar Sastry, Subash Khanal, Adeel Ahmad, Eric Xing, and Nathan Jacobs. Range: Retrieval augmented neural fields for multi-resolution geo-embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24680–24689, June 2025.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, May 2012. ISSN 0034-4257. doi: 10.1016/j.rse.2011.11.026. URL http://dx.doi.org/10.1016/j.rse.2011.11.026.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34818–34831, June 2026.

Peng Gong, Han Liu, Meinan Zhang, Congcong Li, Jie Wang, Huabing Huang, Nicholas Clinton, Luyan Ji, Wenyu Li, Yuqi Bai, Bin Chen, Bing Xu, Zhiliang Zhu, Cui Yuan, Hoi Ping Suen, Jing Guo, Nan Xu, Weijia Li, Yuanyuan Zhao, Jun Yang, Chaoqing Yu, Xi Wang, Haohuan Fu, Le Yu, Iryna Dronova, Fengming Hui, Xiao Cheng, Xueli Shi, Fengjin Xiao, Qiufeng Liu, and Lianchun Song. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Science Bulletin*, 64(6):370–373, March 2019. ISSN 2095-9273. doi: 10.1016/j.scib.2019.03.002. URL http://dx.doi.org/10.1016/j.scib.2019.03.002.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394, October 2025.

C.O. Justice, E. Vermote, J.R.G. Townshend, R. Defries, D.P. Roy, D.K. Hall, V.V. Salomonson, J.L. Privette, G. Riggs, A. Strahler, W. Lucht, R.B. Myneni, Y. Knyazikhin, S.W. Running, R.R. Nemani, Zhengming Wan, A.R. Huete, W. van Leeuwen, R.E. Wolfe, L. Giglio, J. Muller, P. Lewis, and M.J. Barnsley. The moderate resolution imaging spectroradiometer (modis): land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1228–1249, July 1998. ISSN 0196-2892. doi: 10.1109/36.701075. URL http://dx.doi.org/10.1109/36.701075.

Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C. Mazzariello, Mark Mathis, and Steven P. Brumby. Global land use / land cover with sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 4704–4707. IEEE, July 2021. doi: 10.1109/igarss47720.2021.9553499. URL http://dx.doi.org/10.1109/IGARSS47720.2021.9553499.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):4347–4355, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i4.32457. URL http://dx.doi.org/10.1609/aaai.v39i4.32457.

Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the earth. *Nature Ecology & Evolution*, 7(11):1778–1789, Nov 2023. ISSN 2397-334X. doi: 10.1038/s41559-023-02206-6. URL https://doi.org/10.1038/s41559-023-02206-6.

Wenyuan Li, Shunlin Liang, Keyan Chen, Yongzhe Chen, Han Ma, Jianglei Xu, Yichuan Ma, Yuxiang Zhang, Shikang Guan, Husheng Fang, and Zhenwei Shi. Agrifm: A multi-source temporal remote sensing foundation model for agriculture mapping. *Remote Sensing of Environment*, 334:115234, March 2026. ISSN 0034-4257. doi: 10.1016/j.rse.2026.115234. URL http://dx.doi.org/10.1016/j.rse.2026.115234.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *The Eighth International Conference on Learning Representations*. openreview, 2020.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, August 2023. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2023.06.016. URL http://dx.doi.org/10.1016/j.isprsjprs.2023.06.016.

Dimitrios Michail, Eleni Saka, Ioannis Giannopoulos, and Ioannis Papoutsis. Osmgraphclip: Learning global location representations from openstreetmap graphs, 2026. URL https://arxiv.org/abs/2606.08046.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://doi.org/10.1038/323533a0.

Marc Ruß wurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1746–1759, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/073c8584ef86bee26fe9d639ec648e28-Paper-Conference.pdf.

M. Santoro, O. Cartus, N. Carvalhais, D. M. A. Rozendaal, V. Avitabile, A. Araza, S. de Bruin, M. Herold, S. Quegan, P. Rodríguez-Veiga, H. Balzter, J. Carreiras, D. Schepaschenko, M. Korets, M. Shimada, T. Itoh, Á. Moreno Martínez, J. Cavlovic, R. Cazzolla Gatti, P. da Conceição Bispo, N. Dewnath, N. Labrière, J. Liang, J. Lindsell, E. T. A. Mitchard, A. Morel, A. M. Pacheco Pascagaza, C. M. Ryan, F. Slik, G. Vaglio Laurin, H. Verbeeck, A. Wijaya, and S. Willcock. The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. *Earth System Science Data*, 13(8):3927–3950, 2021. doi: 10.5194/essd-13-3927-2021. URL https://essd.copernicus.org/articles/13/3927/2021/.

Craig Saunders, Alex Gammerman, and Vladimir Vovk. *Ridge Regression learning algorithm in dual variables*, pp. 515–521. Morgan Kaufmann, 1998. Edited by J.Shavlik.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Carlos Gomes, Ankur Kumar, Vishal Gaur, Myscon Truong, Denys Godwin, Sam Khallaghi, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Disha Shidham, Rufai Omowunmi Balogun, Venkatesh Kolluru, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Timothy Mayer, Christopher Hain, Robert Kennedy, Bianca Zadrozny, David Bell, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-eo-2.0: A versatile multitemporal

foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. ISSN 1558-0644. doi: 10.1109/tgrs.2025.3642610. URL http://dx.doi.org/10.1109/TGRS.2025.3642610.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60280–60300. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/tseng25a.html.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8690–8701. Curran Associates, Inc., 2023. doi: 10.52202/075280-0379. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b57aaddf85ab01a2445a79c9edc1f4b-Paper-Conference.pdf.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9888–9899, October 2025.

Michael A. Wulder, Thomas R. Loveland, David P. Roy, Christopher J. Crawford, Jeffrey G. Masek, Curtis E. Woodcock, Richard G. Allen, Martha C. Anderson, Alan S. Belward, Warren B. Cohen, John Dwyer, Angela Erb, Feng Gao, Patrick Griffiths, Dennis Helder, Txomin Hermosilla, James D. Hipple, Patrick Hostert, M. Joseph Hughes, Justin Huntington, David M. Johnson, Robert Kennedy, Ayse Kilic, Zhan Li, Leo Lymburner, Joel McCorkel, Nima Pahlevan, Theodore A. Scambos, Crystal Schaaf, John R. Schott, Yongwei Sheng, James Storey, Eric Vermote, James Vogelmann, Joanne C. White, Randolph H. Wynne, and Zhe Zhu. Current status of landsat program, science, and applications. *Remote Sensing of Environment*, 225:127–147, May 2019. ISSN 0034-4257. doi: 10.1016/j.rse.2019.02.015. URL http://dx.doi.org/10.1016/j.rse.2019.02.015.

Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino. Esa worldcover 10 m 2021 v200, October 2022. URL https://doi.org/10.5281/zenodo.7254221.

A Additional Method Details

A.1 Task Loss and Output Projection

Writing $\widehat{y}_g$ for the scalar regression output and $\widehat{p}_g \in \Delta^{K-1}$ for the projected class-probability vector, the risk in Eq. (1) is

$$\mathcal{R}(g) = \mathbb{E}[(y - \widehat{y}_g(q))^2] \qquad \text{for regression,} \tag{12}$$

$$\mathcal{R}(g) = \mathbb{E}\left[\sum_{k=1}^{K}(e_{y,k} - \widehat{p}_{g,k}(q))^2\right] \qquad \text{for classification.} \tag{13}$$

Here Δ^{K-1} is the probability simplex: its vectors have K nonnegative entries summing to one. The latter loss is multiclass Brier risk. Since $e_y \in \Delta^{K-1}$, the nonexpansiveness of Euclidean projection gives, for any $a \in \mathbb{R}^K$,

$$\|e_y - P_K(a)\|_2 = \|P_K(e_y) - P_K(a)\|_2 \le \|e_y - a\|_2. \tag{14}$$

Thus projection cannot increase Brier loss. The theoretical analysis uses the unprojected residual $r = u(y) - f(z)$ and risk

$$\widetilde{\mathcal{R}}(g) = \mathbb{E}\|r - g(q)\|_2^2. \tag{15}$$

A.2 Geographic and Representation Distances

Let $s = (\phi, \lambda)$, with latitude ϕ and longitude λ in radians. Its unit-sphere coordinate and great-circle distance are

$$v(s) = (\cos\phi\cos\lambda, \cos\phi\sin\lambda, \sin\phi), \qquad d_s(s, s') = R_\oplus \arccos(v(s)^\top v(s')), \tag{16}$$

where $R_\oplus$ is Earth's mean radius. Embedding channels are standardized using fitting-data means μ_z and scales σ_z, with unit scale for zero-variance channels. The resulting similarity and angular distance are

$$\begin{aligned} \mathrm{sim}(z, z') &= \frac{[(z - \mu_z) \oslash \sigma_z]^\top [(z' - \mu_z) \oslash \sigma_z]}{\|(z - \mu_z) \oslash \sigma_z\|_2 \|(z' - \mu_z) \oslash \sigma_z\|_2}, \\ d_z(z, z') &= \arccos\big(\mathrm{sim}(z, z')\big), \end{aligned} \tag{17}$$

where $\oslash$ denotes componentwise division. For the distance $d_{k_a}(q)$ to the k_a-th geographically closest candidate, the bandwidth is

$$h(q) = \max\{d_{k_a}(q), h_{\min}\}. \tag{18}$$

The rank $1 \le k_a \le M$ sets the bandwidth, not a hard cutoff on retrieval weights; $h_{\min} > 0$ is the minimum bandwidth.

A.3 Residual Scaling

For a given retrieval configuration, the scale is $\eta_\theta(q) = \mathrm{sigmoid}(\theta_0 + \boldsymbol{\theta}^\top c(q))$, with $0 \le \eta_\theta(q) \le 1$. Here $\theta = (\theta_0, \boldsymbol{\theta})$ contains the intercept and five feature coefficients. The feature vector summarizes the nearest candidate distance $d_1(q) = \min_i d_s(s_q, s_i)$, bandwidth, local density, effective retrieval size, and residual dispersion:

$$\rho(q) = \log k_a - 2\log\max\{h(q), 1\}, \tag{19}$$

$$N_{\mathrm{eff}}(q) = \left[\sum_i w_i(q)^2\right]^{-1}, \tag{20}$$

$$v_r(q) = \frac{1}{K}\sum_i w_i(q)\|r_i - b(q)\|_2^2, \qquad \tau_r = \frac{1}{nK}\sum_{i=1}^{n}\|r_i\|_2^2. \tag{21}$$

Here $v_r(q)$ measures residual dispersion and τ_r is the mean residual energy per output component over the n memory entries. Distances enter the logarithms in kilometers, and $K = 1$ for regression. The five unstandardized features are

$$[\log(1 + d_1(q)), \log(1 + h(q)), \rho(q), \log(1 + N_{\mathrm{eff}}(q)), \log(1 + v_r(q)/\tau_r)]^\top. \tag{22}$$

Their fitting-data means and scales define $c(q)$. The fitted scale is

$$\eta_{\theta^*}(q) = \mathrm{sigmoid}(\theta_0^* + \boldsymbol{\theta}^{*\top} c(q)). \tag{23}$$

The scale is fitted by Eq. (5) to residuals from cross-fitted base predictions, with self-retrieval excluded (Appendix C). The coefficient vector is penalized by $\lambda_\eta/5$, while the intercept is unpenalized. The fitted correction and output are

$$\widehat{g}(q) = \eta_{\theta^*}(q) b(q), \qquad \widehat{u}(q) = P_K(f(z_q) + \widehat{g}(q)). \tag{24}$$

B Theoretical Analysis

B.1 Optimal Residual Augmentation

Throughout the proofs, the fitted base predictor is held fixed and $\mathbb{E}\|r\|_2^2 < \infty$. The function $m(q) = g^*(q) = \mathbb{E}[r \mid q]$ is square-integrable by conditional Jensen's inequality. For any square-integrable correction $g(q)$,

$$r - g(q) = [r - m(q)] + [m(q) - g(q)].$$

The cross term in its squared norm vanishes by the tower property:

$$\begin{aligned} \mathbb{E}\langle r - m(q), m(q) - g(q)\rangle &= \mathbb{E}[\langle \mathbb{E}[r - m(q) \mid q], m(q) - g(q)\rangle] \\ &= 0. \end{aligned} \tag{25}$$

Consequently,

$$\widetilde{\mathcal{R}}(g) = \widetilde{\mathcal{R}}(g^*) + \mathbb{E}\|m(q) - g(q)\|_2^2. \tag{26}$$

The second term is nonnegative and vanishes for $g = g^*$, proving Lemma 1. Setting $g = 0$ gives the recoverable risk

$$\widetilde{\mathcal{R}}(0) - \widetilde{\mathcal{R}}(g^*) = \mathbb{E}\|m(q)\|_2^2. \tag{27}$$

These identities concern unprojected squared loss. For classification, Eq. (14) shows that the final simplex projection cannot increase loss relative to the one-hot target.

B.2 Residual Retrieval Error

Write each memory residual as

$$r_i = m(q_i) + \varepsilon_i, \qquad m(q_i) = \mathbb{E}[r_i \mid q_i]. \tag{28}$$

The fitted base predictor and retrieval geometry are fixed for this conditional calculation. Write $\mathcal{Q} = (q, q_1, \ldots, q_M)$ for the query and its selected candidate queries, including the selection from the memory. The weights are functions of this geometry and are fixed given $\mathcal{Q}$; they are not selected using the noise values being averaged. Thus

$$b(q) = \sum_{i=1}^{M} w_i(q) r_i, \qquad w_i(q) \geq 0, \qquad \sum_{i=1}^{M} w_i(q) = 1. \tag{29}$$

Here i indexes the query's candidates, not all memory entries. Assume the joint smoothness condition

$$\|m(q_i) - m(q)\|_2 \leq L_s d_s(s_i, s) + L_z d_z(z_i, z), \tag{30}$$

where $L_s, L_z \geq 0$, and the distances are defined in Appendix A.2. The required noise conditions, conditional on the same geometry, are

$$\begin{aligned} \mathbb{E}[\varepsilon_i \mid \mathcal{Q}] = 0, \qquad \mathbb{E}[\|\varepsilon_i\|_2^2 \mid \mathcal{Q}] \leq \sigma^2, \\ \mathbb{E}[\langle \varepsilon_i, \varepsilon_j\rangle \mid \mathcal{Q}] = 0 \quad (i \neq j). \end{aligned} \tag{31}$$

The bound controls total noise energy across the K output components.

Using $N_{\mathrm{eff}}(q)$ from Eq. (20), define

$$R_s(q) = \sum_i w_i(q) d_s(s_i, s), \qquad R_z(q) = \sum_i w_i(q) d_z(z_i, z). \tag{32}$$

The effective size is 1 for a single nonzero weight and M for uniform weights.

Substitution gives the bias–noise decomposition in Eq. (11). For its bias term, the triangle inequality and Eq. (30) give

$$\left\| \sum_i w_i(q)[m(q_i) - m(q)] \right\|_2 \le \sum_i w_i(q) \|m(q_i) - m(q)\|_2 \le L_s R_s(q) + L_z R_z(q). \tag{33}$$

The bias is fixed given $\mathcal{Q}$, and the weighted noise has conditional mean zero, so their cross term has zero conditional expectation. Expanding the noise norm gives

$$\mathbb{E}\left[\left\| \sum_i w_i \varepsilon_i \right\|_2^2 \,\middle|\, \mathcal{Q} \right] = \sum_i w_i^2 \mathbb{E}[\|\varepsilon_i\|_2^2 \mid \mathcal{Q}] + \sum_{i \ne j} w_i w_j \mathbb{E}[\langle \varepsilon_i, \varepsilon_j \rangle \mid \mathcal{Q}] \le \sigma^2 \sum_i w_i^2 = \frac{\sigma^2}{N_{\text{eff}}(q)}. \tag{34}$$

Here and below $w_i = w_i(q)$ within the conditional calculation. Combining squared bias and noise proves

$$\mathbb{E}[\|b(q) - m(q)\|_2^2 \mid \mathcal{Q}] \le U(q), \qquad U(q) = [L_s R_s(q) + L_z R_z(q)]^2 + \frac{\sigma^2}{N_{\text{eff}}(q)}. \tag{35}$$

The first term reflects geographic and representation mismatch; the second reflects noise amplification by concentrated weights.

Risk improvement. For a fresh query residual $r_q = m(q) + \varepsilon_q$, assume $\mathbb{E}[\varepsilon_q \mid \mathcal{Q}] = 0$, finite conditional second moment, and conditional independence of ε_q from the memory noise. Expanding the two squared norms yields

$$\begin{aligned} \mathbb{E}[\|r_q\|_2^2 \mid \mathcal{Q}] &= \|m(q)\|_2^2 + \mathbb{E}[\|\varepsilon_q\|_2^2 \mid \mathcal{Q}], \\ \mathbb{E}[\|r_q - b(q)\|_2^2 \mid \mathcal{Q}] &= \mathbb{E}[\|m(q) - b(q)\|_2^2 \mid \mathcal{Q}] + \mathbb{E}[\|\varepsilon_q\|_2^2 \mid \mathcal{Q}]. \end{aligned} \tag{36}$$

The query-noise cross term vanishes under these conditions. Subtracting cancels its energy and gives

$$\mathbb{E}[\|r_q\|_2^2 - \|r_q - b(q)\|_2^2 \mid \mathcal{Q}] = \|m(q)\|_2^2 - \mathbb{E}[\|b(q) - m(q)\|_2^2 \mid \mathcal{Q}]. \tag{37}$$

Therefore

$$U(q) < \|m(q)\|_2^2 \tag{38}$$

is sufficient, not necessary, for positive conditional risk reduction. For classification, the identity applies before projection; the projection argument is given in Appendix A.1. The risk identity describes the unscaled retrieval estimate $b(q)$.

C Experimental Setup

C.1 Data, Targets, and Inputs

The experiments use 7 learned embeddings and 3 raw sensor inputs at 7,140 globally distributed locations. The classification targets are ESRI LULC, ESA WorldCover, and FROM-GLC10 (Karra et al., 2021; Zanaga et al., 2022; Gong et al., 2019); the regression targets are canopy height, above-ground biomass, and Copernicus elevation (Lang et al., 2023; Santoro et al., 2021; cop, 2022) (Fig. 8). Target-specific missing labels are removed within each fold. Classification uses the majority valid class in a centered 5×5 window, and regression uses the mean valid value in a centered 3×3 window.

Satellite Inputs. Each spatial tensor is mean-pooled to a vector per observation before fitting a downstream predictor. The learned embeddings range from 13 to 1,024 dimensions, whereas the sensor OBS. inputs contain 6 or 7 channels (Table 5). Each source is evaluated independently, without combining satellite sources.

Table 5: Input dimensions and sources.

Input	Type	D	Source/Venue	Year
AlphaEarth (Brown et al., 2025)	Emb.	64	GEE	2025
TESSERA (Feng et al., 2026)	Emb.	128	CVPR	2026
Copernicus-FM (Wang et al., 2025)	Emb.	768	ICCV	2025
SDC (Chen et al., 2026)	Emb.	13	ESSD	2026
AgriFM (Li et al., 2026)	Emb.	1,024	RSE	2026
Prithvi (Szwarcman et al., 2026)	Emb.	768	NASA NTRS	2024
TerraMind (Jakubik et al., 2025)	Emb.	384	ICCV	2025
Sentinel-2 (Drusch et al., 2012)	Obs.	6	GEE	-
Landsat (Wulder et al., 2019)	Obs.	6	GEE	-
MODIS (Justice et al., 1998)	Obs.	7	GEE	-

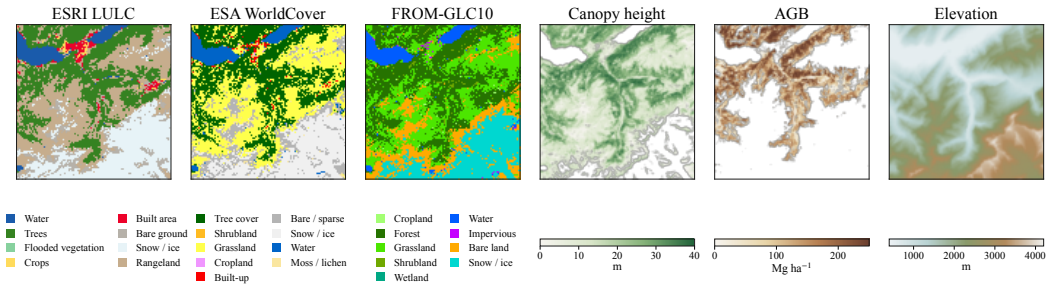


Figure 8: Examples of the 6 downstream targets.

C.2 Data Splits and Fitting Protocol

Each input is paired with Linear, random forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), CatBoost (Prokhorenkova et al., 2018), MLP (Rumelhart et al., 1986), and RBF kernel ridge (Saunders et al., 1998). Five mutually exclusive IID folds are used in turn for testing. Within each outer training set, cross-fitted base predictions produce the residual memory before the final predictor is refitted on the full outer training set. Test labels do not enter model fitting or configuration selection. Predictor settings are given in Appendix C.4.

C.3 Comparison Methods and Metrics

All methods use the same base predictions, cross-fitted residuals, and outer folds. The comparison includes XYZ, Direct, SH-L15, Wrap (Mac Aodha et al., 2019), Space2Vec-G/T (Mai et al., 2020), SphereC/C+ and SphereM/M+ (Mai et al., 2023), GeoCLIP (Vivanco Cepeda et al., 2023), SatCLIP (Klemmer et al., 2025), RANGE/RANGE+ (Dhakal et al., 2025), TTE (Cher et al., 2026), and OSMGC (Michail et al., 2026). Classification models are selected by Brier risk and reported with OA and F1; regression models are selected by MSE and reported with R^2 and RMSE. Encoder features, preprocessing, and checkpoints are specified in Appendix C.5.

C.4 Predictor Settings

Linear models use logistic regression with $C = 1$ and ridge regression with $\alpha = 1$. Random forests use 300 trees, maximum depth 16, minimum leaf size 5, and square-root feature subsampling. XGBoost uses 300 trees, depth 6, learning rate 0.05, row subsampling 0.85, and feature subsampling 0.85. CatBoost uses 400 iterations, depth 8, learning rate 0.05, and ℓ_2 leaf penalty 3. The MLP has hidden widths 256 and 128, GELU activations, dropout 0.1, and 100 AdamW epochs. RBF kernel ridge uses $\alpha = 1$ and $\gamma = 1/D$. These final settings were chosen through grid search for good performance across most model–dataset combinations.

Median imputation is fitted inside each fold. Linear, MLP, and kernel models standardize inputs; tree models use imputed values directly. The final outer-fold fits use seeds $42, 43, 44, 45, 46$; inner cross-fitting uses deterministic seeds derived from the outer fold and inner partition.

SATRAP uses a candidate-pool size of 32, bandwidth rank 8, semantic coefficient 4, minimum bandwidth 25 km, and scaling penalty 10^{-4}. The candidate-pool size, bandwidth rank, and semantic coefficient are selected by squared prediction loss on predictions for the training observations.

C.5 LOCATION-ENCODING SETTINGS

Each encoder maps a geographic coordinate $s = (\phi, \lambda)$, with latitude ϕ and longitude λ, to a vector $e(s) \in \mathbb{R}^{D_e}$. Feature generation uses coordinates only; downstream labels and satellite inputs are not passed to the frozen encoder in these experiments. Analytic encoders have no fitted weights. Pretrained encoders are loaded from the indicated checkpoints, placed in evaluation mode, and kept frozen.

Table 6: Location encoders, publication sources, and experimental settings.

Family	Encoder	D_e	Venue	Year	Configuration
Coordinate	DIRECT	2	-	-	Longitude and latitude in radians, using the constant offset of the reference implementation.
	XYZ	3	-	-	Unit-sphere Cartesian coordinates.
	WRAP	4	-	-	Sine and cosine of longitude and latitude.
	SH-L15	255	-	-	Real spherical harmonics of degrees 1 through 15; the constant degree-zero term is omitted.
Multiscale	SPACE2VEC-T (Mai et al., 2020)	192	ICLR	2020	Theory encoder; $F = 32$, radius range $[1, 360]$, geometric frequencies, and 3 axes separated by 120°.
	SPACE2VEC-G (Mai et al., 2020)	256	ICLR	2020	Grid encoder; $F = 64$, radius range $[1, 360]$, geometric frequencies.
	SPHEREC (Mai et al., 2023)	384	ISPRS JPRS	2023	SPHEREC encoder; $F = 64$, radius range $[1, 360]$.
	SPHERECP (Mai et al., 2023)	768	ISPRS JPRS	2023	SPHEREC with the additional Grid terms; $F = 64$, radius range $[1, 360]$.
	SPHEREM (Mai et al., 2023)	480	ISPRS JPRS	2023	SPHEREM encoder; $F = 48$, radius range $[10, 360]$.
	SPHEREMP (Mai et al., 2023)	512	ISPRS JPRS	2023	SPHEREM with its additional single-coordinate terms; $F = 32$, radius range $[10, 360]$.
Pretrained	GEOCLIP† (Vivanco Cepeda et al., 2023)	512	NeurIPS	2023	Official pretrained GEOCLIP location encoder.
	SATCLIP† (Klemmer et al., 2025)	256	AAAI	2025	`SatCLIP-ResNet18-L10` checkpoint.
	RANGE† (Dhakal et al., 2025)	1280	CVPR	2025	RANGE with the SatCLIP-ViT16-L40 checkpoint, medium retrieval database, and RANGE's retrieval coefficient $\beta = 0.5$.
	RANGE+† (Dhakal et al., 2025)	1280	CVPR	2025	RANGE+ with the same checkpoint, database, and retrieval coefficient β.
	TTE† (Cher et al., 2026)	512	ECCV	2026	Official `MVRL/TTE` pretrained encoder.
	OSMGC† (Michail et al., 2026)	256	arXiv	2026	OSMGraphCLIP-A-L10 checkpoint.

† denotes frozen pretrained weights; F is the number of spatial frequencies.

Direct Coordinate Encodings. Longitude λ and latitude ϕ follow the radian convention in Appendix A.2. The unit-sphere encoding is

$$e_{\mathrm{XYZ}}(s) = v(s), \tag{39}$$

which is continuous across the antimeridian and does not duplicate the poles. WRAP uses

$$e_{\mathrm{Wrap}}(s) = (\cos\lambda, \sin\lambda, \cos\phi, \sin\phi). \tag{40}$$

DIRECT follows the reference implementation and applies the same constant $-\pi$ offset after converting both coordinate columns to radians. Because the residual head includes an intercept and every feature is centered with training-fold statistics, this constant translation has no effect on fitted predictions. XYZ and WRAP follow the implementations evaluated by Ruß wurm et al. (2024); WRAP was introduced as a geographic prior by Mac Aodha et al. (2019).

The SH-L15 feature vector is

$$e_{\mathrm{SH}}(s) = [Y_{\ell m}(s)]_{\ell=1,\ldots,15;\ m=-\ell,\ldots,\ell}\,, \tag{41}$$

where $Y_{\ell m}$ denotes a real spherical harmonic. Negative orders use the imaginary component of the complex harmonic, positive orders use its real component, and nonzero orders receive the conventional $\sqrt{2}$ factor and phase sign. Omitting $\ell = 0$ gives $\sum_{\ell=1}^{15}(2\ell + 1) = 255$ channels. The maximum degree is fixed at 15 for every split, input, task, and predictor.

SPACE2VEC and SPHERE2VEC. We use the analytic positional-encoding layers from the reference implementations of SPACE2VEC and SPHERE2VEC (Mai et al., 2020; 2023). For F frequencies and radius limits $r_{\min}$ and $r_{\max}$, the timescales and frequencies are

$$\tau_j = r_{\min} \exp\left(\frac{j}{F-1} \log \frac{r_{\max}}{r_{\min}}\right), \qquad \omega_j = \tau_j^{-1}, \qquad j = 0, \ldots, F-1. \tag{42}$$

The coordinates supplied to these encoders are ordered as longitude followed by latitude and retain their degree units, matching the radius parameters in Table 6.

SPACE2VEC-G applies sine and cosine at every frequency independently to both coordinate axes, producing $4F$ channels. SPACE2VEC-T projects the coordinate onto the 3 unit vectors $(1, 0)$, $(-1/2, \sqrt{3}/2)$, and $(-1/2, -\sqrt{3}/2)$, then applies sine and cosine at each frequency, producing $6F$ channels. SPHEREC replaces independent Cartesian products with spherical terms based on scaled latitude, scaled longitude, and their trigonometric products. SPHERECC+ additionally includes the corresponding independent Grid terms. SPHEREM combines scaled and unscaled latitude–longitude factors; SPHEREM+ adds the independent sine and cosine components. The variant-specific frequency counts and radius limits in Table 6 are fixed before downstream fitting.

Pretrained encoders. GEOCLIP uses its official location encoder with pretrained weights enabled (Vivanco Cepeda et al., 2023). SATCLIP loads `microsoft/SatCLIP-ResNet18-L10` (Klemmer et al., 2025). RANGE and RANGE+ load the SatCLIP-ViT16-L40 location checkpoint together with the authors' medium retrieval database and use RANGE's retrieval coefficient $\beta = 0.5$ (Dhakal et al., 2025). TTE loads `MVRL/TTE` (Cher et al., 2026), and OSMGC loads the OSMGraphCLIP-A-L10 location encoder (Michail et al., 2026). GEOCLIP and TTE receive coordinates in latitude–longitude order; SATCLIP, RANGE, RANGE+, and OSMGC receive longitude–latitude order, following each encoder's coordinate convention.

Use within SATRAP. For location-embedding retrieval (Table 4), distances between encoded locations replace great-circle distance while the base predictor, residual memory, semantic similarity, and scaling function remain fixed. Encoder features are standardized with training-fold statistics; RANGE and RANGE+ use their train-fitted min–max transform. Candidate entries are obtained by Euclidean distance in the encoded space. To place those distances on the same numerical scale as the geodesic reference, each fold multiplies them by a training-only constant so that the median distance to the k_ath candidate matches the corresponding median great-circle distance. The location-encoder comparison uses $M = 32$, $k_a = 8$, $\beta = 4$, minimum bandwidth 25 km, and scaling penalty 10^{-4} for every encoder.

D GEOGRAPHIC INFORMATION ANALYSIS

D.1 GEOGRAPHIC ATTRIBUTE PREDICTION

We examine whether each satellite input contains information about absolute location and related geographic attributes. Geolocation predicts unit-sphere XYZ coordinates and is evaluated by median great-circle error. Additional probes predict Köppen–Geiger climate class, country, continent, distance to the coast, BIO1 temperature, and BIO12 precipitation. LINEAR and random-forest probes are fitted on the predefined training split and evaluated on the validation split.

All 10 inputs support geolocation substantially better than predicting the mean training location. The strongest embedding reaches a median error of 767 km, compared with 6,673 km for the mean-location baseline. Geographic classes are also recoverable: the best embedding obtains macro-F1 values of 0.701 for Köppen–Geiger climate, 0.610 for country, and 0.980 for continent. Environmental variables remain predictable, with best R^2 values of 0.625 for distance to the coast, 0.979 for BIO1 temperature, and 0.773 for BIO12 precipitation (Fig. 1, left).

D.2 ITERATIVE LINEAR LOCATION REMOVAL

To distinguish linearly and nonlinearly accessible location structure, we iteratively remove the embedding components used by a linear location predictor. Let N be the number of observations used

to fit that predictor, and let $\mathbf{Z}_0 \in \mathbb{R}^{N\times D}$ contain their embeddings after channel-wise standardization with training statistics. Their unit-sphere locations form

$$\mathbf{S} = \begin{bmatrix} v(s_1)^\top \\ \vdots \\ v(s_N)^\top \end{bmatrix} \in \mathbb{R}^{N\times 3}. \tag{43}$$

At iteration t, a multi-output ridge regression predicts the 3 location coordinates from the current representation:

$$\mathbf{B}_t = \arg\min_{\mathbf{B}\in\mathbb{R}^{D\times 3}} \|\mathbf{Z}_{t-1}\mathbf{B} - \mathbf{S}\|_F^2 + \alpha\|\mathbf{B}\|_F^2, \tag{44}$$

where $\mathbf{B}_t \in \mathbb{R}^{D\times 3}$ is the fitted coefficient matrix, $\alpha > 0$ is the ridge coefficient, and $\|\cdot\|_F$ is the Frobenius norm. Its compact singular value decomposition is $\mathbf{B}_t = \mathbf{U}_t\Sigma_t\mathbf{V}_t^\top$. The columns of $\mathbf{U}_t \in \mathbb{R}^{D\times \mathrm{rank}(\mathbf{B}_t)}$ are the embedding-space directions used for linear location prediction, Σ_t contains their singular values, $\mathbf{V}_t$ contains the corresponding directions in the 3-dimensional location target, and $\mathrm{rank}(\mathbf{B}_t) \le 3$.

Let $\mathbf{Q}_{t-1}$ contain all previously selected orthonormal directions, with $\mathbf{Q}_0$ empty. We remove from $\mathbf{U}_t$ the component already represented by $\mathbf{Q}_{t-1}$ and retain an orthonormal basis for the nonzero remainder:

$$\mathbf{C}_t = (I_D - \mathbf{Q}_{t-1}\mathbf{Q}_{t-1}^\top)\mathbf{U}_t, \qquad \mathbf{A}_t = \mathrm{orth}(\mathbf{C}_t), \qquad \mathbf{Q}_t = [\,\mathbf{Q}_{t-1}\ \mathbf{A}_t\,]. \tag{45}$$

Here I_D is the $D \times D$ identity matrix, $\mathbf{A}_t$ contains only newly identified directions, and orth returns orthonormal columns spanning the nonzero column space of its argument. Every iteration projects the original standardized matrix:

$$\mathbf{Z}_t = \mathbf{Z}_0(I_D - \mathbf{Q}_t\mathbf{Q}_t^\top). \tag{46}$$

The embedding matrix retains its $N \times D$ dimensions, while its effective rank is reduced by the number of columns in $\mathbf{Q}_t$. A new location predictor is fitted on $\mathbf{Z}_t$ at the next iteration.

For a dimensionality-matched control, let $\mathbf{Q}_t^{\mathrm{rand}}$ be a fixed, nested random orthonormal basis with the same cumulative rank as $\mathbf{Q}_t$. We use

$$\mathbf{Z}_t^{\mathrm{rand}} = \mathbf{Z}_0\big(I_D - \mathbf{Q}_t^{\mathrm{rand}}(\mathbf{Q}_t^{\mathrm{rand}})^\top\big), \qquad \mathrm{rank}(\mathbf{Q}_t^{\mathrm{rand}}) = \mathrm{rank}(\mathbf{Q}_t). \tag{47}$$

Each condition runs for at least 1 and at most 10 rounds. Evaluation stops after a completed round when linear median geolocation error reaches 95% of the mean-location baseline, moves toward that baseline by no more than 100 km in 2 consecutive rounds, or no new linearly independent direction remains. LINEAR and random-forest geolocation models are refitted after every projection.

The iterative projection sharply degrades the linear location predictor for most inputs, verifying removal of the directions identified at each round. The RF probe is substantially more robust: in Fig. 1 (right, panel a), the solid circular trajectories show the LINEAR probe and the dashed square trajectories show the RF probe. After the final projection, ALPHAEARTH, COPERNICUS-FM, PRITHVI, and TESSERA (Brown et al., 2025; Wang et al., 2025; Szwarcman et al., 2026; Feng et al., 2026) retain median RF errors of 1,785, 1,507, 1,962, and 3,118 km, respectively, all below the 6,673-km mean-location baseline.

For learned embeddings, downstream changes after removing the selected linear location directions are small and close to those caused by equal-rank random removal. In Fig. 1 (right, panels b and c), points near the diagonal indicate comparable changes under the 2 interventions, while points below the diagonal indicate a larger decrease after removing the location-selected directions. Mean LINEAR classification F1 changes from 0.581 to 0.573 after location-component removal and to 0.572 after random removal. Mean LINEAR regression R^2 changes from 0.505 to 0.488 and 0.489, respectively. The intervention characterizes downstream dependence on the embedding subspace selected by successive linear geolocation predictors.

Geography-only prediction. Coordinates, Köppen class, country, continent, coast distance, and WorldClim variables predict all 6 downstream targets without imagery. These results establish the predictive value of geography.

E Additional Experimental Results

E.1 Residual Structure and Retrieval

For every test query, we compute $R_s(q)$, $R_z(q)$, $N_{\text{eff}}(q)$, and the reduction in squared error before scaling and projection. Each input–task–predictor–fold combination is grouped into equal-count quantiles independently for the three retrieval quantities. The analysis uses directly observed residuals and retrieval quantities across all test queries.

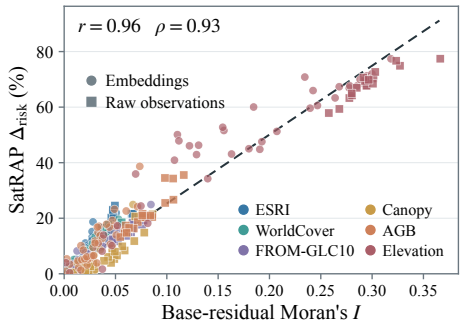


Figure 9: Base-residual geographic structure and SatRAP Δ_{risk}.

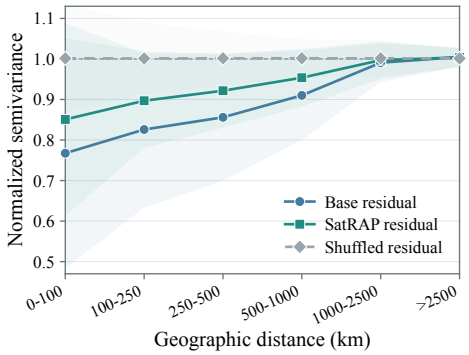


Figure 10: Geographic dependence before and after SatRAP.

Across input–task–predictor combinations, base-residual Moran's I is strongly associated with SatRAP Δ_{risk} (Pearson $r = 0.96$, Spearman $\rho = 0.93$) in Fig. 9. This association is consistent with a recoverable residual component $m(q)$ in Lemma 1.

In the grouped classification results, realized gain decreases from the lowest to highest groups of both R_s and R_z. Larger N_{eff} is more consistently associated with stability; cross-fold variation decreases across effective-size groups for both task families. Regression means are not uniformly monotone because broader weights can reduce variance while increasing geographic or representation mismatch. These empirical relationships are consistent with the bias–variance trade-off described in Proposition 1.

After augmentation, short-range semivariance moves toward the shuffled-residual reference and mean Moran's I falls from 0.058 to 0.008 (Fig. 10). Thus, SatRAP reduces both risk and the geographic dependence remaining in the base errors.

E.2 Parameter Sensitivity

Fig. 11 separates classification and regression performance as candidate-pool size, bandwidth rank, and semantic weight vary. The two task families exhibit similar responses. A very small candidate pool limits the improvement; enlarging it initially raises both OA and R^2, but further expansion gradually reduces the gain. Increasing the bandwidth rank likewise weakens performance after the initial peak. These trends favor selective retrieval over indiscriminately averaging a larger set of residuals. Semantic weighting has a different effect: both metrics improve as embedding agreement receives greater emphasis, then decline when the weights become too concentrated. This behavior is consistent with Proposition 1: broader retrieval can reduce noise through a larger effective size while increasing geographic or represen-

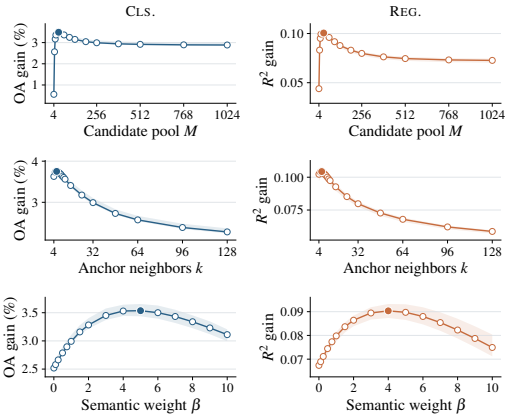


Figure 11: Parameter analysis.

tation mismatch. Strong semantic weighting reverses this trade-off, reducing representation mismatch but also effective size (Fig. 7). The main-experiment settings are given in Appendix C.4.

E.3 Results by Input and Base Predictor

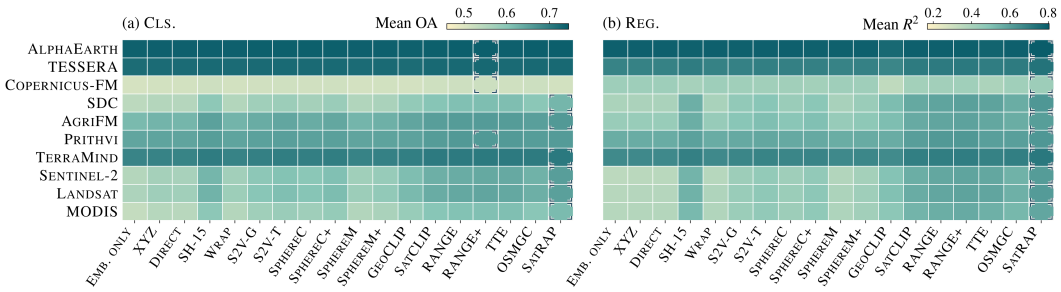


Figure 12: Mean overall accuracy (left) and mean R^2 (right). ⁚ denotes the best method for each input.

Fig. 12 shows how method rankings depend on the satellite input. RANGE+ (Dhakal et al., 2025) attains the highest mean OA with AlphaEarth and TESSERA, whereas SatRAP leads with SDC, AgriFM, TerraMind, and all three sensor inputs. Regression shows a more consistent ordering: SatRAP has the highest mean R^2 for every input, with RANGE+ generally the closest competitor. For TESSERA, the corresponding values are 0.732 and 0.714; for SDC, they are 0.624 and 0.599.

Table 7: Mean test Δ_{risk} by method and base predictor (%). Values are mean ± standard deviation over five splits; each split averages ten inputs and six tasks.

Method	Linear	RF	XGBoost	CatBoost	MLP	Kernel ridge
XYZ	0.6 ± 0.1	0.4 ± 0.1	0.8 ± 0.1	1.9 ± 0.1	1.8 ± 0.1	0.5 ± 0.1
Direct	0.6 ± 0.1	0.3 ± 0.1	0.7 ± 0.1	1.8 ± 0.1	1.7 ± 0.1	0.4 ± 0.1
SH-L15	10.5 ± 1.1	12.3 ± 1.4	8.5 ± 1.0	11.1 ± 1.2	8.1 ± 0.9	9.5 ± 1.3
Wrap (Mac Aodha et al., 2019)	1.1 ± 0.1	1.0 ± 0.2	1.2 ± 0.2	2.3 ± 0.2	2.1 ± 0.2	0.9 ± 0.1
Space2Vec-G (Mai et al., 2020)	4.7 ± 0.8	5.2 ± 1.1	3.9 ± 0.7	5.7 ± 0.9	4.6 ± 0.6	4.0 ± 1.0
Space2Vec-T (Mai et al., 2020)	6.4 ± 0.5	7.3 ± 0.6	5.3 ± 0.5	7.3 ± 0.5	5.6 ± 0.4	5.7 ± 0.5
SphereC (Mai et al., 2023)	4.3 ± 0.5	4.7 ± 0.5	3.7 ± 0.4	5.3 ± 0.6	4.5 ± 0.4	3.9 ± 0.4
SphereC+ (Mai et al., 2023)	6.3 ± 1.0	7.3 ± 1.0	5.0 ± 0.7	7.1 ± 0.9	5.5 ± 0.6	5.5 ± 1.0
SphereM (Mai et al., 2023)	1.1 ± 0.2	1.1 ± 0.2	1.1 ± 0.1	2.4 ± 0.1	2.2 ± 0.1	0.9 ± 0.2
SphereM+ (Mai et al., 2023)	3.3 ± 0.5	3.7 ± 0.7	3.0 ± 0.5	4.5 ± 0.6	3.8 ± 0.4	3.0 ± 0.7
GeoCLIP† (Vivanco Cepeda et al., 2023)	4.4 ± 1.8	6.6 ± 1.1	2.8 ± 0.8	5.3 ± 1.0	2.9 ± 0.8	3.5 ± 1.1
SatCLIP† (Klemmer et al., 2025)	11.9 ± 1.1	13.7 ± 1.2	9.5 ± 0.9	12.3 ± 1.0	9.0 ± 0.7	10.7 ± 1.2
RANGE† (Dhakal et al., 2025)	13.5 ± 1.1	15.5 ± 1.3	10.6 ± 1.0	13.7 ± 1.2	10.0 ± 0.8	12.2 ± 1.3
RANGE+† (Dhakal et al., 2025)	$\underline{14.5 \pm 1.1}$	$\underline{16.6 \pm 1.2}$	$\underline{11.5 \pm 1.0}$	$\underline{14.6 \pm 1.1}$	$\underline{10.7 \pm 0.7}$	$\underline{13.2 \pm 1.3}$
TTE† (Cher et al., 2026)	12.3 ± 1.5	14.4 ± 1.7	9.1 ± 1.2	12.3 ± 1.4	8.4 ± 1.0	10.6 ± 1.7
OSMGC† (Michail et al., 2026)	11.4 ± 1.1	13.1 ± 1.2	9.2 ± 0.9	12.0 ± 1.0	8.9 ± 0.7	10.3 ± 1.2
SatRAP (Ours)	$\mathbf{18.0 \pm 0.6}$	$\mathbf{19.6 \pm 0.8}$	$\mathbf{12.5 \pm 0.5}$	$\mathbf{16.2 \pm 0.5}$	$\mathbf{11.4 \pm 0.4}$	$\mathbf{15.2 \pm 0.7}$

Bold and underline mark the best and second-best value within each predictor. † denotes a pretrained location encoder.

Table 7 averages risk reduction over ten inputs and six tasks for each predictor, with standard deviations across splits. SatRAP achieves the largest mean reduction for every predictor, followed by RANGE+. The reductions are 18.0% versus 14.5% for Linear, 19.6% versus 16.6% for RF, and 15.2% versus 13.2% for Kernel ridge. SatRAP also outperforms RANGE+ with XGBoost, CatBoost, and MLP. SatRAP also has lower cross-split standard deviations than RANGE+ for every predictor.

The other methods retain similar rankings across predictors. XYZ, Direct, and Wrap (Mac Aodha et al., 2019) yield modest reductions; SH-L15 is substantially stronger. Space2Vec-T exceeds Space2Vec-G (Mai et al., 2020), and the SphereC+ and SphereM+ variants improve on their respective counterparts (Mai et al., 2023). SatCLIP (Klemmer et al., 2025) and OSMGC (Michail et al., 2026) outperform GeoCLIP (Vivanco Cepeda et al., 2023) for every predictor. TTE (Cher et al., 2026) is competitive with these methods, while RANGE and RANGE+ achieve larger reductions. The advantage of SatRAP thus persists across linear, tree-based, neural, and kernel predictors.