CHARACTERIZING BRAZILIAN ATLANTIC FOREST RESTORATION OUTCOMES WITH GEOSPATIAL ALPHA-EARTH EMBEDDINGS

Alice Heiman
Department of Computer Science
Stanford University
{aheiman}@stanford.edu

ABSTRACT

The Atlantic Forest in Brazil is a critical biodiversity hotspot, yet less than 12% of its original cover remains. While monitoring forest restoration at scale is essential, traditional methods are limited by the impracticality of large-scale on-the-ground reporting and by the saturation of remote-sensing indices such as NDVI. We study 3,929 restoration sites using satellite embeddings from the Alpha-Earth foundation model to evaluate their utility for predicting early restoration success. We introduce a "Reference Trajectory Embedding", defining a novel success metric based on cosine similarity to persistent forest reference sites. Using 5-fold spatial cross-validation, we show that incorporating foundation model embeddings and baseline similarity improves over pure environmental factors in predicting future restoration trajectories. Our results suggest that, with additional work, embeddings may be used to monitor and quantify restoration trajectories and identify sites following recovery paths similar to those of known reference forests. Finally, we introduce an open-source geospatial package for polygon-first trajectory extraction to support the scaling of these analyses (https://github.com/aliceheiman/gee-polygons).

1 INTRODUCTION

Brazil, short for "Terra do Brasil" (eng. land of brazilwood), is home to some of the world's most important and biodiverse forests. Yet, you won't find the tree that gave Brazil its name in the Amazon Rainforest. The now endangered tree, paubrasila, can only be found in the Atlantic Forest. The Atlantic Forest provides ecosystem services to over 150 million Brazilians, contains over 20,000 plant species, and supplies 62% of Brazil's hydroelectric power (de Lima et al., 2020; World Wildlife Fund, 2025). However, less than 12% of its original forest area remains, making it a World Restoration Flagship project as designated by the United Nations (UN) (UN Decade on Restoration, 2025). Initiatives like the Trinational Atlantic Forest Pact have attracted over 300 organizations to restore the forest. Yet, planting alone is insufficient. Effective restoration requires rigorous monitoring to understand what drives ecological success. However, traditional vegetation monitoring metrics, such as NDVI, often saturate in dense tropical regions and lack universal thresholds for success (Gao et al., 2020).

Earth observation foundation models (FMs), such as Alpha-Earth, produce learned embeddings that combine multiple data types. These high-dimensional vectors may encode more ecological information than simplistic indices. In this paper, we propose using pre-trained satellite embeddings to characterize and predict forest restoration outcomes. Our contributions include: (1) a novel "Reference Trajectory Embedding" success metric using cosine similarity to persistent forest reference sites; (2) a systematic evaluation on 3,929 restoration polygons showing that embeddings outperform environmental baselines in restoration trajectory prediction tasks; and (3) an open-source geospatial package for polygon-first trajectory analysis.[1]

[1] https://github.com/aliceheiman/gee-polygons

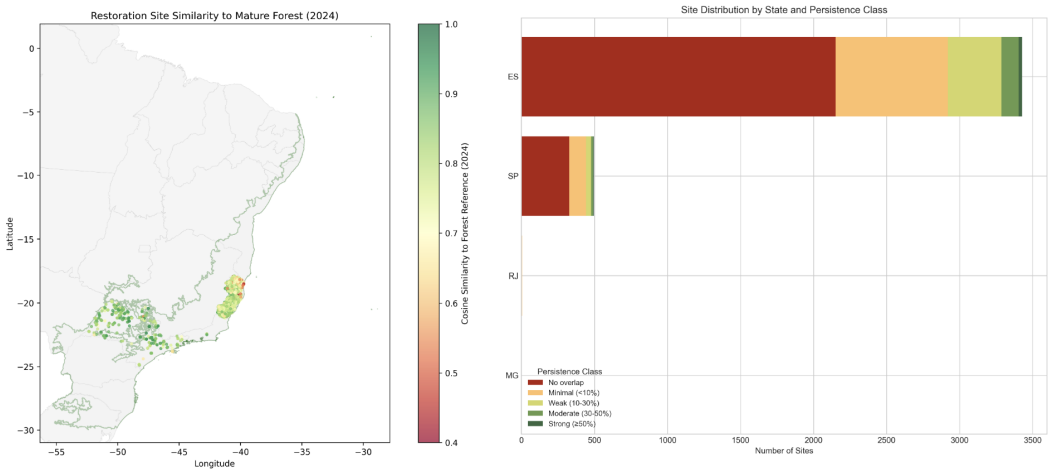

Figure 1: Study area in the Atlantic Forest biome, Brazil. (left): Restoration sites colored by their similarity to the mean mature secondary forest. (right): Restoration site distribution by state, colored by the percentage overlap with persistent secondary forest in 2024.

2 Data and Methods

2.1 Study Sites

We use 3,929 restoration polygons (mean area 2.87 ha) from the Observatorio da Restauração e Reflorestamento (ORR). The ORR dataset originally contained 40,476 polygons from the Atlantic Forest, which we filtered to projects with areas of 1 hectare or more and a start year of 2021 (see Appendix). We chose 2021 as a common start year to streamline trajectory analyses, and since 2021 was the year with the most restoration project starts.

2.2 Feature Extraction

All features were extracted via Google Earth Engine at the polygon level.

- **AlphaEarth embeddings**: 64 dimensions per site per year (2019-2024), extracted as polygon-mean values. We use 2020, 2021, and 2024 embeddings as pre-treatment, restoration start, and post-treatment embeddings, respectively.
- **Environmental**: annual precipitation, temperature (min/max), elevation, slope, aspect, distance to nearest forest, forest proportion within 1 km.
- **NDVI**: Sentinel-2 annual NDVI mean, amplitude, and variance. Project metadata: area, shape complexity, and restoration strategy as given in the ORR dataset.

See the Appendix for a full breakdown of data sources used.

2.3 Success Metrics

We extract a reference set of 1,000 persistent secondary forest points using the MapBiomas Collection 10 Land Use Land Cover classification, focusing on pixels that were forested for at least 10 years until 2024. Additionally, we identify 101 negative reference points from urban, pasture, and agricultural areas for comparison and extract features for each. Next, we calculate the cosine similarity between each site's 2024 embedding and the mean embedding of the 1,000 forest points, using a threshold of 0.8411 (90th percentile of negative reference similarity) to determine similarity to persistent secondary forests. We also match each restoration site with its closest persistent secondary forest pixel for comparison and assess the overlap of MapBiomas secondary forest regrowth from 2021 to 2024 with restoration sites. This allows us to compute three success metrics.

- **Global Forest Similarity**: Similarity of restoration site at 2024 to the mean of mature secondary forest.
- **Local Forest Similarity**: Similarity of restoration site at 2024 to the closest mature secondary forest pixel.
- **Persistent Forest Overlap**: Percentage of restoration site at 2024 classified as persistent forest with growth start 2021-2024.

In practice, a high Global Forest Similarity score indicates that a site's spectral and structural properties have converged toward those of mature secondary forest, providing a remotely sensed proxy for the restoration trajectory without requiring field visits.

3 Experimental Design

We pose restoration outcome prediction as three regression tasks: 1) percentage persistent forest, 2) cosine similarity to the global reference site, and 3) cosine similarity to the local reference site. We employ 5-fold spatial cross-validation, using k-means clustering on coordinates to ensure folds are geographically separated. We compare Ridge Regression, Random Forest (100 trees), and XGBoost (100 estimators) across several feature sets.

Table 1: Feature sets used in ablation experiments.

Feature Set	**Dims**	**Description**
Project	4	Area, shape, method, strategy
Env	9	Climate, topography, landscape
NDVI	5	Pre-treatment NDVI stats
Emb_{pre}	64	Pre-treatment embeddings (2020)
$\text{Emb}_{\text{start}}$	64	Start-year embeddings (2021)
Emb_{Δ}	64	Post $-$ pre embeddings (2022)
Sim_{T_0}	1	Similarity pre-treatment embeddings and reference site
Project+Env+NDVI	18	Classical baseline
Emb+Env	73	Key hypothesis test

4 Results

We observe that embeddings significantly improve performance for similarity-based success metrics. However, the improvement from incorporating embeddings to predict the percentage of persistent forest is smaller. This is likely due to the distance to the nearest forest, which acts as an important confounding factor. We also see that predicting future local similarity does not generalize across spatial folds, indicating the need to incorporate features that account for baseline similarity to the reference site.

When analyzing the SHAP feature importances (Table 3), we see that the foundation model embeddings contribute the majority of the predictive signal (55.9%), followed by environmental variables (37.4%). Among individual features, the distance to existing forest fragments is the most significant predictor, in line with previous studies on forest persistence. Interestingly, while NDVI (mean 2020) remains in the top-10 of feature importances, its overall categorical contribution (4.2%) is marginal compared to the geospatial embeddings. This suggests that Alpha-Earth embeddings incorporate spectral greenness while providing additional context that NDVI alone cannot capture.

5 Discussion

Our experiments indicate that satellite embeddings could improve the characterization of forest restoration trajectories. A key finding is the necessity of baseline similarity anchoring. Without knowing the initial similarity to the reference site, environmental features and restoration embeddings fail to generalize across spatial folds ($R^2 < 0$ for Local Similarity), including the baseline

Table 2: Spatial 5-fold Cross-Validation R^2 (mean $\pm$ std), comparison between feature sets.

Feature Set	% Persistent	Sim. Centroid	Sim. Local
Without baseline similarity			
Env (Only)	0.232 ± 0.273	0.039 ± 0.150	-0.316 ± 0.596
NDVI (Only)	0.036 ± 0.176	-0.128 ± 0.346	-0.455 ± 1.024
Emb_{pre} (Only)	0.174 ± 0.089	0.545 ± 0.157	-0.080 ± 0.504
Emb + Env	$\mathbf{0.312 \pm 0.117}$	0.553 ± 0.103	-0.169 ± 0.617
With local baseline similarity at T_0			
Sim_{T_0} (Only)	-0.442 ± 0.914	$\mathbf{0.599 \pm 0.092}$	0.633 ± 0.294
Sim_{T_0} + Env	0.297 ± 0.240	0.578 ± 0.106	0.628 ± 0.259
Sim_{T_0} + All	0.275 ± 0.402	0.548 ± 0.103	$\mathbf{0.651 \pm 0.232}$

Table 3: Feature importance by category (XGB, persistent_success target).

Category	N Features	Total Mean SHAP	% Total
Embedding	64	6.575	55.9%
Environmental	9	4.392	37.4%
NDVI	5	0.488	4.2%
Project	4	0.303	2.6%

similarity results in reasonable predictions ($R^2 = 0.65$). However, we acknowledge that the strong predictive signal from Sim_{T_0} may partially reflect temporal autocorrelation since sites that are already similar to the reference forest at baseline are likely to remain so. Disentangling these effects will require longer time series and matched non-restored control sites. Also, the Reference Trajectory Embedding does not implement a formal treatment-control comparison or difference-in-differences estimator as in, for instance, BACI designs.

Additionally, Alpha-Earth is a proprietary model, and it remains an open question whether the observed gains generalize to open foundation models such as Clay or SatMAE. Benchmarking across architectures is an important direction for future work. This study focused on the Alpha-Earth model within the Atlantic Forest biome. Future work should evaluate the transferability of these findings to other biomes (e.g., the Amazon) and across different foundation model architectures. Additionally, the short time window (2021-2024) only captures the initial stages of restoration. As the temporal availability of embeddings increases, it will be exciting to confirm if these early signals correlate with longer-term forest recovery. In the Appendix, we include a brief exploration into using embeddings for unsupervised clustering analysis. Without any prior project knowledge, the restoration projects form distinctive clusters. We are excited to continue work on using embeddings to find insightful relationships between restoration projects and to help characterize projects with missing data. Moreover, a similar similarity approach could also be used to identify degraded regions.

We note that while our method shows potential, it is not yet mature enough to, for instance, automatically punish or reject restoration sites. Moreover, although remote-sensing methods can be useful, they should be seen as complements rather than replacements for community engagement.

6 Conclusion

We propose a novel framework for using earth observation foundation models to characterize forest restoration projects in the Brazilian Atlantic Forest. Our "Reference Trajectory Embedding" is based on using embedding cosine similarities between on-going restorations and known mature reference forests to quantify and track restoration trajectories. Our results show that initial conditions largely influence future trajectories. We hope our open-source package facilitates the analysis of polygons and their trajectories. We look forward to exploring the utility of embeddings further with our Brazilian collaborators.

ACKNOWLEDGMENTS

We thank the Observatorio da Restauração e Reflorestamento (ORR) for providing restoration polygon data. Thank you also to Hilary Brumberg and Professor Ahmed Ragab for giving valuable thoughts and insights.

Use of LLMs: We used Claude Code to help develop the code for Google Earth Engine scripts, experimental scripts, and data exploration and figures notebooks. We also used Google Gemini to help format tables and references in LaTeX. We used Google Gemini and Grammarly to identify spelling mistakes and improve writing in this paper.

REFERENCES

Renato AF de Lima, Alexandre A Oliveira, Guilherme R Pitta, André L de Gasper, Alexander C Vibrans, Jérôme Chave, Hans Ter Steege, and Paulo I Prado. The erosion of biodiversity and biomass in the Atlantic Forest biodiversity hotspot. *Nature Communications*, 11(1):6347, 2020. doi: 10.1038/s41467-020-19909-2.

Thiago C Dias, Luís F Silveira, ZI Pironkova, and Mercival R Francisco. Greening and browning trends in a tropical forest hotspot: Accounting for fragment size and vegetation indices. *Remote Sensing Applications: Society and Environment*, 26:100751, 2022. doi: 10.1016/j.rsase.2022.100751.

Lianpeng Gao, Xihan Wang, Brian A Johnson, Qingjiu Tian, Yanan Wang, Jochem Verrelst, Xihan Mu, and Xingfa Gu. Remote sensing algorithms for estimation of fractional vegetation cover using pure vegetation index values: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:364–377, 2020. doi: 10.1016/j.isprsjprs.2019.11.018.

UN Decade on Restoration. Trinational Atlantic Forest Pact. `http://www.decadeonrestoration.org/trinational-atlantic-forest-pact`, 2025. Accessed: 2025-08-18.

World Wildlife Fund. Atlantic Forest. `https://www.worldwildlife.org/places/atlantic-forest`, 2025. Accessed: 2026-02-06.

A Appendix

A.1 Appendix A: Study Sites and Data Sources

This section details the geographic distribution of the 3,929 restoration sites and the multi-modal datasets used for feature extraction.

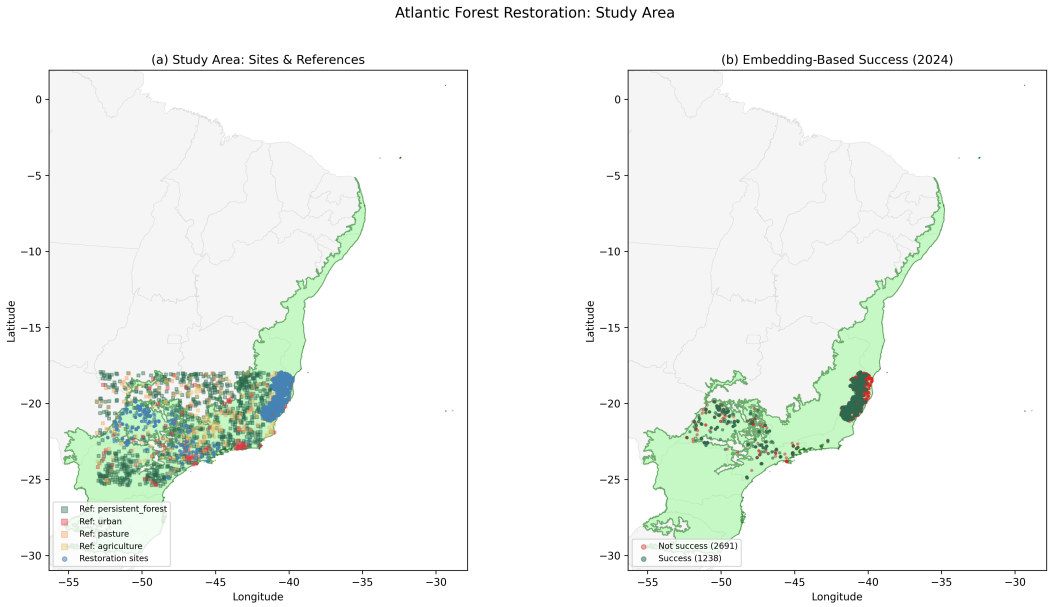

Figure 2: Study area in the Atlantic Forest biome, Brazil. (a) Restoration sites (blue) and reference points. (b) Sites colored by embedding-based success in 2024.

Table 4: Data sources and derived variables used in this study.

Variable	**Data Source**	**Details**
NDVI	COPERNICUS/S2_SR_HARMONIZED	Annual composites 2019–2024; cloud-masked.
LULC Coverage	MapBiomas Brazil Collection 10	Remapped to 6 primary ecological categories.
Persistent Forest	Custom MapBiomas Layer	Age and start year of secondary vegetation regrowth.
Topography	USGS/SRTMGL1_003	Elevation, slope, aspect; 30 m resolution.
Climate	IDAHO_EPSCOR/TERRACLIMATE	Annual precip, temp (max/min), evapotranspiration.
AlphaEarth Embs	GOOGLE/SATELLITE_EMBEDDING/V1	64-dim annual composites (2019–2024).

A.2 Appendix B: Similarity Distributions of Reference Sites

We visualize the embedding space and similarity distributions to validate the ”Success” thresholds used in our metrics.

A.3 Appendix C: NDVI-Based Success Metrics

We utilize the Theil-Sen slope of NDVI series and a Mann-Kendall significance test to create a five-class trajectory (2019–2024) (Dias et al., 2022). While NDVI greening is evident in 55.4% of sites,

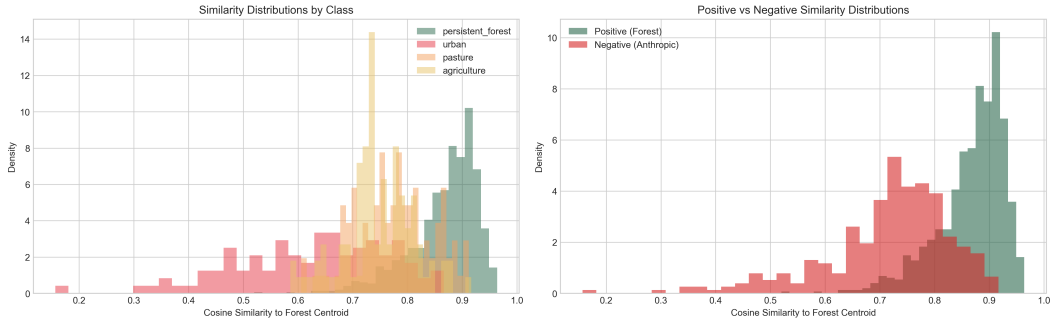

Figure 3: Similarity distributions of reference points to the centroid mature secondary forest.

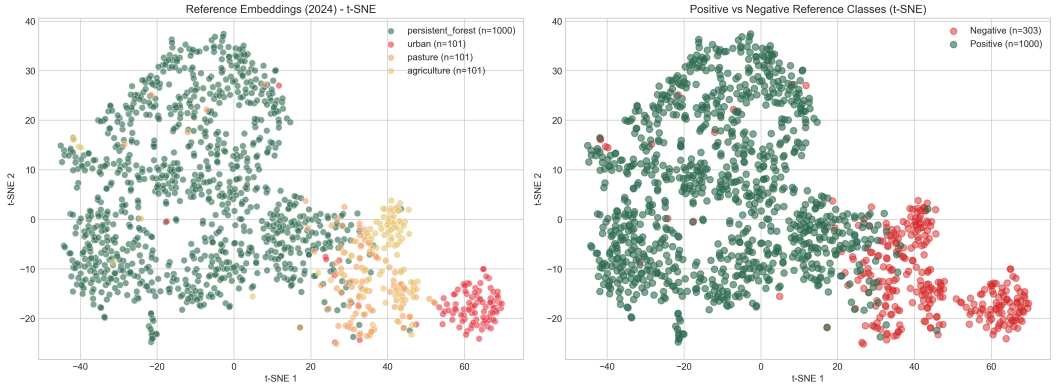

Figure 4: Reference embeddings projected into 2D-space using t-SNE.

it shows poor agreement with structural persistence labels, likely due to saturation and the presence of non-forest greening.

Table 5: NDVI trend-based success tiers derived from Theil-Sen slope estimation.

Success Tier	**Theil-Sen Slope (Annual)**	**Interpretation**
Strong recovery	≥ 0.0090	Rapid greening
Moderate recovery	0.0045–0.0090	Consistent greening
Slight recovery	0.0009–0.0045	Marginal greening
Stable/Stalled	-0.0009 to 0.0009	Unchanged
Degradation	< -0.0009	Browning

A.4 Appendix D: Full Results and Model Ablations

This section provides a complete breakdown of model performance across all architectures and target metrics under 5-fold spatial cross-validation.

A.5 Appendix E: Feature Importance and SHAP Analysis

We use SHAP values to interpret the contribution of different variables to the prediction of restoration success.

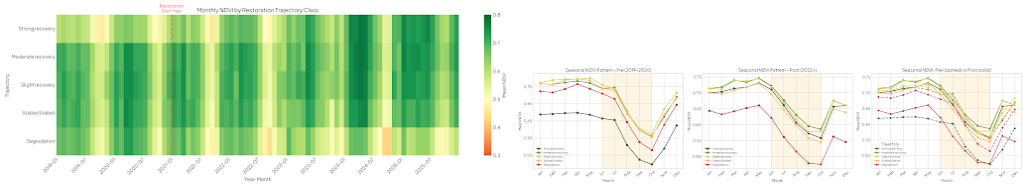

Figure 5: Left: Correlation heatmap of success metrics. Right: Distribution of NDVI slope values.

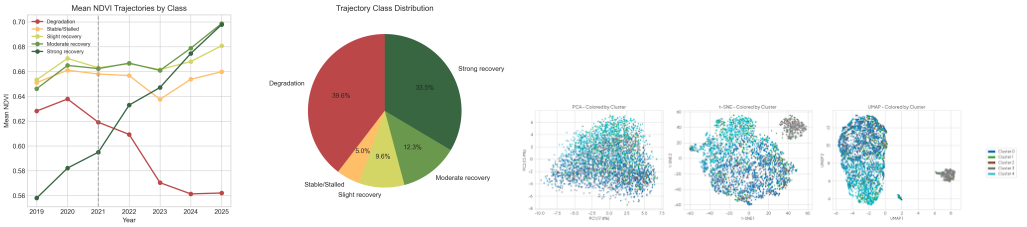

Figure 6: Left: Summary of NDVI success tiers. Right: Spatial clustering of sites based on NDVI profiles.

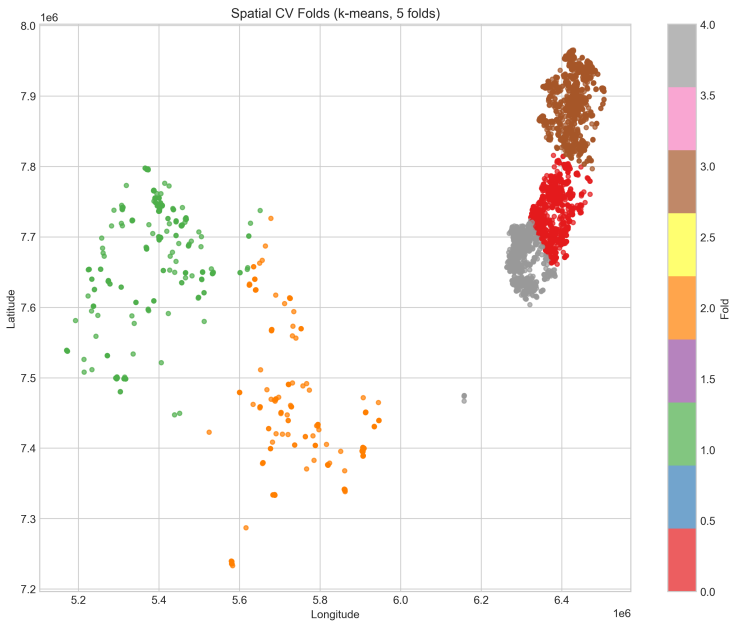

Figure 7: Spatial CV folds for restoration sites generated via K-Means clustering.

Table 6: Comprehensive Spatial CV R^2 scores comparing static and trajectory-based models.

Feature Set	**Target Metric**	**Ridge**	**RF**	**XGB**	**MLP**
Env (Only)	% Persistent	-0.142	0.078	0.093	0.232
NDVI (Only)	% Persistent	-0.076	-0.035	-0.314	0.036
Emb_{pre} + Env	% Persistent	-0.062	0.199	0.241	0.312
Sim_{T_0} (Only)	Sim. Local	0.630	0.591	0.608	0.633
Sim_{T_0} + Emb + Env	Sim. Local	**0.651**	0.575	0.523	-1.358
Sim_{T_0} (Only)	Sim. Centroid	**0.599**	0.566	0.576	0.594

Table 7: 5-fold Cross-Validation Performance Comparison: Best Model Per Feature Set.

	% Persistent (R^2)		Sim. Centroid (R^2)		Sim. Local (R^2)	
Feature Set	Random	Spatial	Random	Spatial	Random	Spatial
Env	0.558	0.232	0.306	0.039	0.293	-0.316
NDVI	0.175	0.036	0.222	-0.128	0.072	-0.455
Emb_{pre}	0.351	0.174	0.698	0.545	0.333	-0.080
Emb+Env	0.567	0.312	0.704	0.553	0.378	-0.169

Table 8: Feature importance by category (XGB, persistent_success target).

Category	**N Features**	**Total Mean SHAP**	**% Total**
Embedding	64	6.575	55.9%
Environmental	9	4.392	37.4%
NDVI	5	0.488	4.2%
Project	4	0.303	2.6%

Table 9: Top 10 individual features by mean SHAP importance.

Feature	**Category**	**Mean SHAP**
dist_to_any_forest	Environmental	2.7560
dist_to_primary_veg	Environmental	0.7187
forest_prop_1km	Environmental	0.5048
pre_A35 (Emb)	Embedding	0.2693
pre_A05 (Emb)	Embedding	0.2432
ndvi_mean_2020	NDVI	0.2405

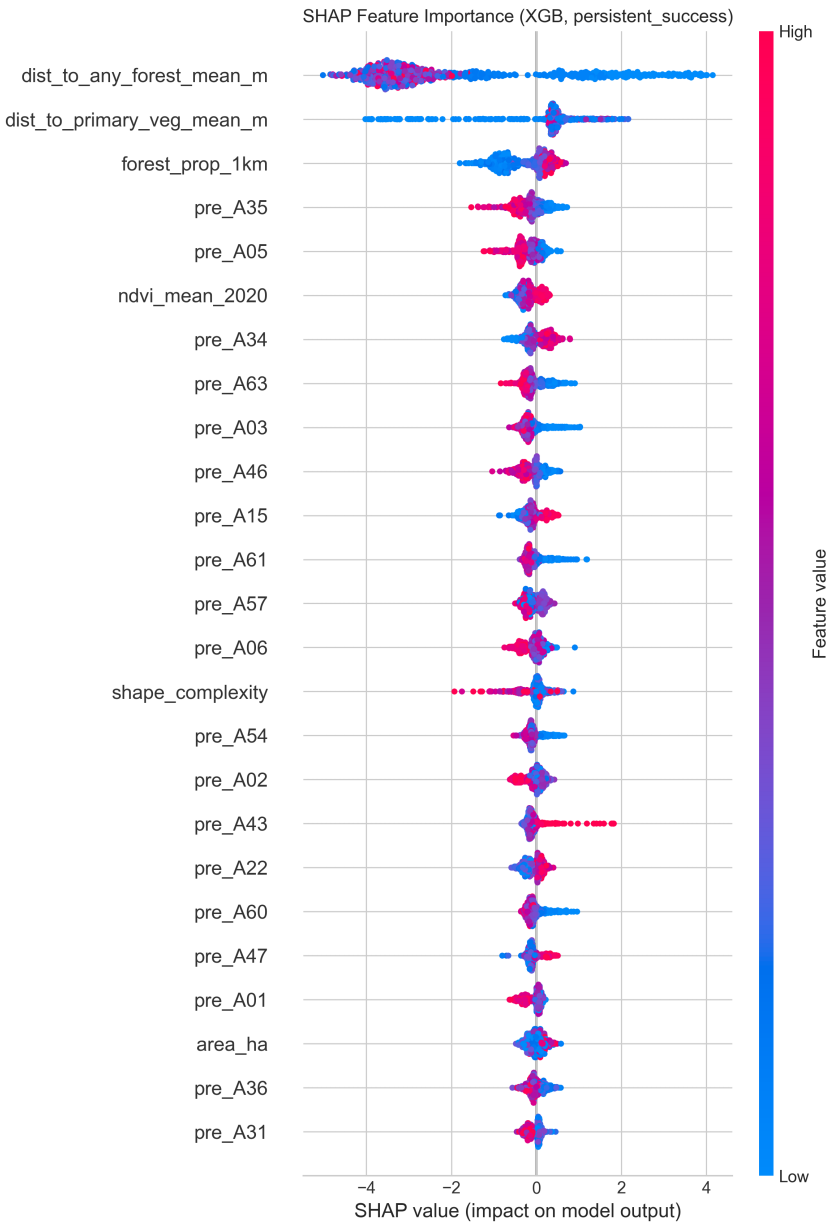

Figure 8: Global SHAP importance summary showing the dominance of spatial context and pre-treatment embeddings.