LABEL-EFFICIENT CHANGE DETECTION WITH PRECOMPUTED SATELLITE EMBEDDINGS UNDER SPATIAL AND TEMPORAL SHIFT

Preetam Chhimpa
Indian Institute of Technology Roorkee
preetamchhimpa2022@gmail.com

ABSTRACT

Label scarcity is a practical bottleneck for satellite change detection. Models must generalize across geography and time with limited human supervision. Using precomputed 64-D annual satellite embeddings from Google Earth Engine, we study Brazil forest-loss detection from MapBiomas land-cover transitions (2017–2022) under tile-level spatial holdout and a held-out future transition (2020→2021). A simple linear probe on embeddings is already strong (F1 $\approx$ 0.98). We then simulate pool-based active learning and show that uncertainty sampling reaches 99.5% of full-data performance with substantially fewer labels than random selection: 3.73×–10.68× fewer labels when random reaches the target, and lower bounds up to >19× when it does not. These results support a practical label-efficient workflow in this proxy-label setting: precomputed embeddings + linear probe + uncertainty-based querying.

1 INTRODUCTION

Precomputed Earth-observation embeddings reduce the need to train large models (Brown et al., 2025), but deployment remains constrained by expensive labeling and failures under geographic and temporal shift (Lu et al., 2004). This work asks a deployment-first question: *given strong pretrained embeddings, how quickly can we reach near full-data performance with limited labels, and which querying strategy is most label-efficient under shift?* We answer this by evaluating a classifier on precomputed annual embeddings for forest-loss detection in Brazil and simulating pool-based active learning under both spatial and temporal settings.

2 DATA AND TASK

We use annual precomputed 64-D satellite embeddings from Google Earth Engine (Gorelick et al., 2017) and MapBiomas Brazil annual land-cover maps (2016–2022) (Souza et al., 2020). For each year-pair $(t{-}1 \rightarrow t)$ we define a binary label: forest loss if a location changes from forest to non-forest, otherwise no loss. Within a Brazil ROI we sample balanced points for five transitions (2017→2018 to 2021→2022) and export a CSV dataset with coordinates, tile id, years, the current embedding, and the embedding difference ($E_t - E_{t-1}$), yielding 100,000 labeled examples. Because MapBiomas labels are produced by an existing remote-sensing mapping pipeline rather than field-verified ground truth, our results should be interpreted as label-efficient learning of an operational proxy target, not as independent ecological validation.

3 METHOD

For each year-pair, we compare two feature choices to predict forest loss: (i) the current-year embedding alone (E_t, 64-D), and (ii) the current embedding augmented with an explicit change feature by concatenation ($[E_t; \Delta_t]$, 128-D), where $\Delta_t := E_t - E_{t-1}$. Although the label is defined over the transition $(t{-}1 \rightarrow t)$, the current-year embedding E_t can already encode the post-change land-cover state and its surrounding context. In particular, recently deforested locations may occupy a distinct

region of embedding space at time t even without explicitly comparing to $t-1$. We therefore compare E_t against $[E_t; \Delta_t]$ to test whether explicit temporal differencing provides additional predictive signal beyond the current representation alone. We fit a linear classifier (logistic regression) and report F1. To evaluate robustness, we use two holdout settings: *spatial holdout* splits the data so that all samples from a given embedding tile appear in only one split (train or test), while *temporal holdout* holds out an unseen year transition (2020→2021) for testing while training on the remaining transitions.

We then simulate pool-based active learning on the training pool: initialize with 5 labeled samples per class and, for 40 rounds, query 100 additional labels using either random selection or uncertainty sampling (Settles, 2009; Lewis & Gale, 1994). For binary logistic regression, let $p_\theta(x) = P_\theta(y{=}1 \mid x)$ denote the current model probability of forest loss. We define uncertainty by distance to the decision boundary,

$$u(x) = -\left|p_\theta(x) - 0.5\right|,$$

and query the 100 unlabeled points with highest $u(x)$, i.e., those whose predicted probabilities are closest to 0.5. At evaluation time, we convert probabilities to class labels using the default 0.5 threshold. We average over three seeds and summarize label efficiency with learning curves, AULC@1000 (mean F1 for label counts ≤ 1000), and the number of labels required to reach 99–99.5% of the full-data baseline.

4 RESULTS

Split	Feature	Full F1	Labels @99.5% (Rand / Uncert)	× Gain
Spatial	E_t	0.9824	410 / 110	3.73
Spatial	$[E_t; \Delta_t]$	0.9831	>4010 / 210†	>19.10
Temporal	E_t	0.9782	1210 / 210	5.76
Temporal	$[E_t; \Delta_t]$	0.9757	3310 / 310	10.68

Table 1: **Label efficiency under spatial and temporal shift.** Labels @99.5% denotes the number of labeled points needed to reach 99.5% of the full-data F1 (random vs. uncertainty sampling). × Gain is the label savings factor (Random / Uncertainty). †Random acquisition did not reach the 99.5% target within the maximum budget of 4010 labels.

Full-data performance. Linear probes on precomputed embeddings are strong (Full F1 ≈ 0.98; Table 1). Spatial holdout achieves 0.9824 (E_t) and 0.9831 ($[E_t; \Delta_t]$); temporal holdout (2020→2021) achieves 0.9782 (E_t) and 0.9757 ($[E_t; \Delta_t]$).

Label efficiency. Uncertainty sampling improves AULC@1000 in all settings (uncertainty vs random): spatial E_t 0.9807 vs 0.9745, spatial $[E_t; \Delta_t]$ 0.9774 vs 0.9704, temporal E_t 0.9751 vs 0.9669, temporal $[E_t; \Delta_t]$ 0.9658 vs 0.9562. At 99.5% of the full-data baseline (Table 1), uncertainty needs 110 vs 410 labels (3.73×) for spatial E_t, 210 vs 1210 (5.76×) for temporal E_t, and 310 vs 3310 (10.68×) for temporal $[E_t; \Delta_t]$; for spatial $[E_t; \Delta_t]$, random does not reach 99.5% within 4010 labels (>19.10× lower bound). At 99.9% of the full-data baseline, random fails within 4010 labels in all settings, while uncertainty reaches the target with 210/510 labels (spatial $E_t/[E_t; \Delta_t]$) and 310/410 (temporal). The full learning curves are shown in Figure 1.

5 CONCLUSION

We studied label-efficient forest-loss detection using precomputed annual satellite embeddings and a simple linear probe under both spatial and temporal shift. Across all settings, full-data linear probes were strong (F1 $\approx$ 0.98), while uncertainty-based active learning substantially reduced annotation requirements relative to random selection. Uncertainty reached 99.5% of the full-data baseline with 3.73×–10.68× fewer labels when random reached the target, and with lower bounds up to >19× otherwise. In this proxy-label setting, these results support a practical workflow: start from strong pretrained EO embeddings, train a lightweight classifier, and prioritize labels using uncertainty-based querying.

We also clarify the scope of this result. Because targets are derived from MapBiomas rather than field-verified ground truth, this study should be interpreted as label-efficient learning of an oper-

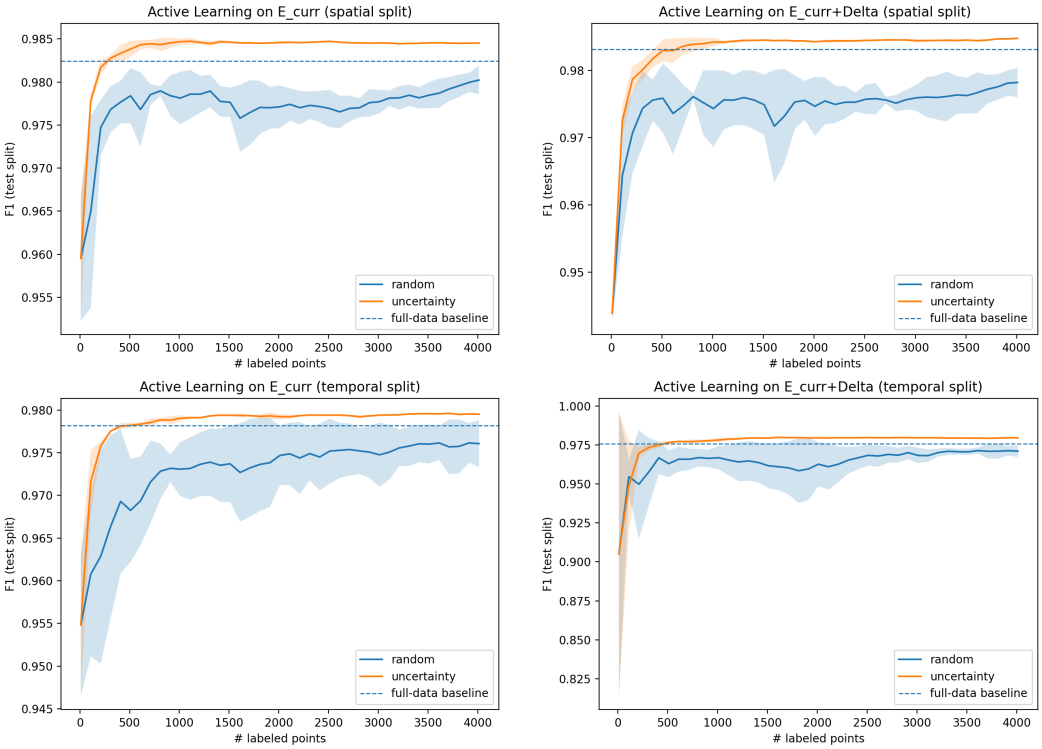

Figure 1: **Active learning curves (mean ± std over 3 seeds).** F1 vs. labeled points; dashed line is the full-data probe. Top: spatial holdout (E_t left, $[E_t; \Delta_t]$ right). Bottom: temporal holdout.

ational proxy target, not as independent ecological validation. In addition, while the main experiments use balanced sampling for controlled comparison, the appendix examines uncertainty behavior around the 0.5 margin, decision-boundary diagnostics, and evaluation under lower positive prevalence. Under imbalance, precision decreases as prevalence decreases, so PR-oriented metrics are more informative for deployment-facing interpretation. Overall, the practical gain here appears to come less from classifier complexity and more from pairing strong pretrained representations with an efficient labeling strategy.

REFERENCES

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. arXiv preprint.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. doi: 10.1016/j.rse.2017.06.031.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994. doi: 10.1007/978-1-4471-2099-5_1.

Dengsheng Lu, Paul Mausel, Eduardo Brondízio, and Emilio Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. doi: 10.1080/0143116031000139863.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009. Computer Sciences Technical Report.

Carlos M. Souza, Julia Z. Shimbo, Marcos R. Rosa, Leandro L. Parente, Ane Alencar, Bernardo F. T. Rudorff, Heinrich Hasenack, Marcelo Matsumoto, Lauro G. Ferreira, Pedro W. M. Souza-Filho, Sebastião W. de Oliveira, Welligton F. Rocha, Alberto V. Fonseca, Camila B. Marques, Cláudio G. Diniz, Dayse Costa, Dayane Monteiro, Enner H. Rosa, Erica Velez-Martin, Eduardo J. Weber, Flávio E. B. Lenti, Felipe F. Paternost, Frans G. C. Pareyn, Júlio V. Siqueira, José L. Viera, Luciano C. F. Neto, Marcelo M. Saraiva, Matheus H. Sales, Mario P. G. Salgado, Roberto Vasconcelos, Sebastião Galano, Valdinei V. Mesquita, and Tasso Azevedo. Reconstructing three decades of land use and land cover changes in brazil with landsat archive and earth engine. *Remote Sensing*, 12(17):2735, 2020. doi: 10.3390/rs12172735.

A Appendix

A.1 Uncertainty Behavior and Threshold Sensitivity

To validate the uncertainty-query rule used in Section 3, we compared predicted-probability distributions on the unlabeled pool at the first acquisition step across three query seeds (0,1,2). Summary statistics in this subsection are averaged over these three seeds. Across all four split/feature settings, queried samples are concentrated near the decision boundary ($p_\theta(y{=}1 \mid x) \approx 0.5$): the fraction in $[0.45, 0.55]$ is enriched by 84.34× on average relative to the unlabeled pool, and mean $|p - 0.5|$ decreases from 0.438 (all unlabeled points) to 0.003 (queried points). This is consistent with margin-based uncertainty sampling for binary logistic regression and is visualized in Figure 2.

We also swept evaluation thresholds for full-data probes. The largest observed F1 improvement over the default threshold 0.5 is 0.003, indicating that a fixed 0.5 threshold is a stable, simple operating point in this setup (Figure 3).

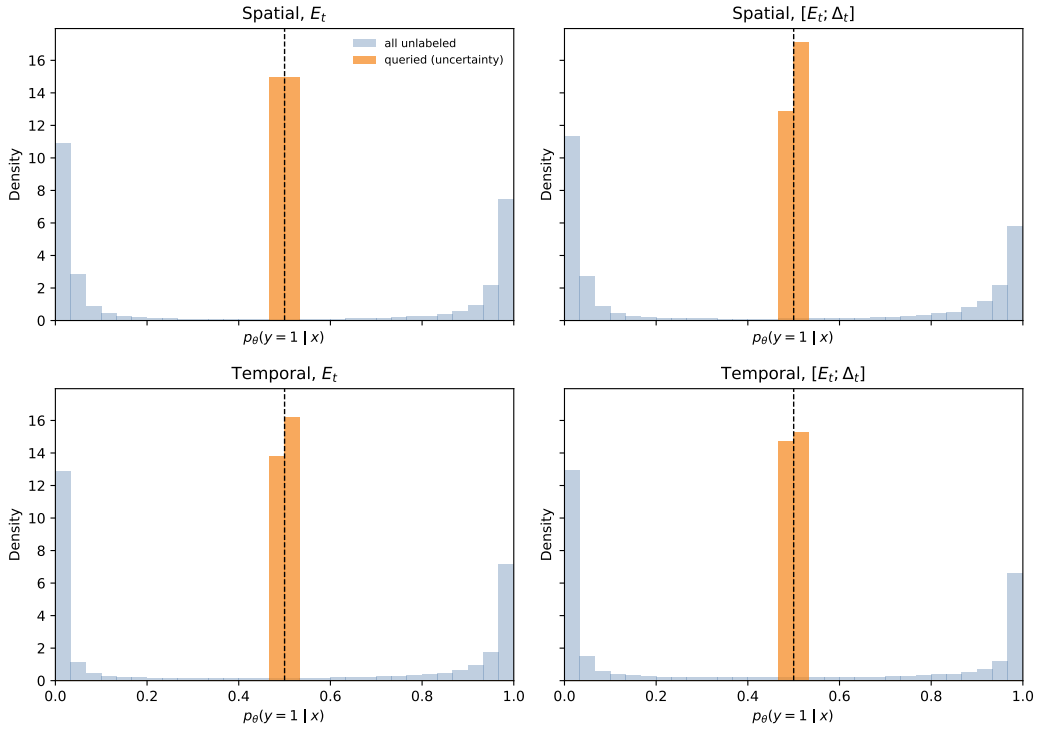

Figure 2: **Uncertainty histogram check (representative seed).** Orange bars are uncertainty-selected queried samples; blue bars are all unlabeled pool samples. Histograms are normalized to density rather than raw counts. Queried points concentrate near probability 0.5 in all settings.

A.2 Imbalanced Evaluation

The main paper uses balanced sampling to isolate label-efficiency effects, but deployment prevalence is typically much lower. To assess this gap, we kept training unchanged and resampled held-out test splits to target positive prevalences of 10%, 5%, and 1% (20 repeats). As expected, precision at threshold 0.5 drops as prevalence decreases, while PR-AUC remains informative for ranking quality (Table 2).

A.3 Decision-Boundary Comparison (Diagnostic)

For each split/feature setting, we trained (i) a full-data probe and (ii) an uncertainty-trained probe at the 99.5% label-efficiency target budget from Table 1 (110, 210, 210, 310 labels), averaging

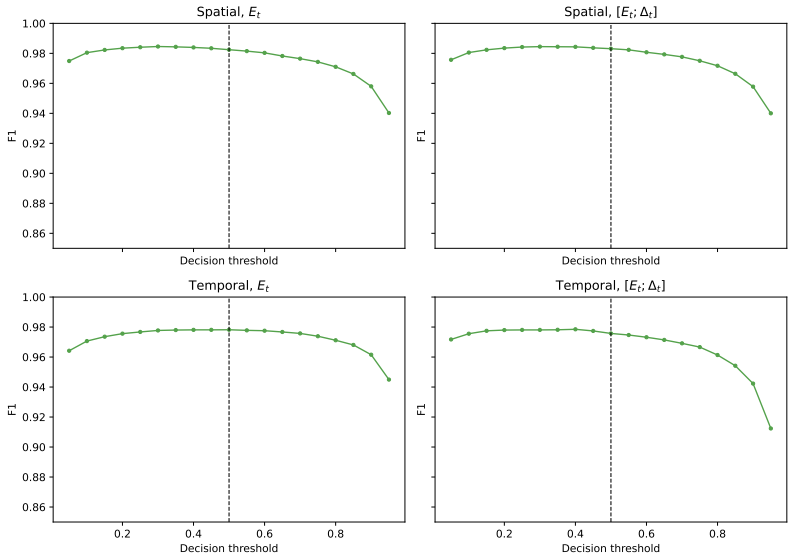

Figure 3: **Threshold sensitivity on held-out splits.** F1 vs. decision threshold for full-data probes. Dashed lines mark threshold 0.5.

Split	Feature	Model	PR-AUC@10%	PR-AUC@5%	PR-AUC@1%	Prec@0.5@10%	Prec@0.5@5%	Prec@0.5@1%
Spatial	E_t	Full-data	0.924	0.857	0.573	0.775	0.619	0.239
Spatial	E_t	Uncertainty@99.5%	0.909	0.833	0.536	0.710	0.537	0.183
Spatial	$[E_t; \Delta_t]$	Full-data	0.917	0.846	0.540	0.773	0.617	0.237
Spatial	$[E_t; \Delta_t]$	Uncertainty@99.5%	0.906	0.827	0.513	0.750	0.587	0.215
Temporal	E_t	Full-data	0.941	0.885	0.633	0.781	0.628	0.245
Temporal	E_t	Uncertainty@99.5%	0.929	0.865	0.579	0.785	0.633	0.249
Temporal	$[E_t; \Delta_t]$	Full-data	0.945	0.893	0.655	0.820	0.683	0.293
Temporal	$[E_t; \Delta_t]$	Uncertainty@99.5%	0.934	0.873	0.603	0.741	0.575	0.207

Table 2: **Imbalanced held-out evaluation with fixed training setup.** Test sets are resampled to lower positive prevalence (10%, 5%, 1%).

Split	Feature	Labels	CosSim(w)	$\|\hat{w}_f - \hat{w}_a\|_2$	$\|\Delta b\|$	Top10 Overlap
Spatial	E_t	110	0.271	1.207	19.434	0.07
Spatial	$[E_t; \Delta_t]$	210	0.256	1.219	14.924	0.13
Temporal	E_t	210	0.416	1.077	22.458	0.37
Temporal	$[E_t; \Delta_t]$	310	0.330	1.156	25.496	0.10

Table 3: **Decision-boundary diagnostic: full-data vs uncertainty-trained probes at 99.5% budgets (mean over 3 query seeds).** CosSim is cosine similarity of original-space coefficient vectors; $\|\hat{w}_f - \hat{w}_a\|_2$ is the L2 distance between normalized coefficients; $|\Delta b|$ is absolute intercept difference; Top10 Overlap is the overlap fraction of top-10 absolute coefficients.

the diagnostic metrics over three query seeds (0,1,2). We then compared linear coefficients in the original feature space (after undoing standardization in the pipeline). Coefficient-space alignment is limited, while predictive performance remains near full-data at these budgets (uncertainty/full-data F1 ratio 0.995–0.998). Detailed diagnostics are reported in Table 3. We therefore treat this as a diagnostic result rather than a central claim.