



# Better Baselines

Foundation Model Embeddings for Credible Conservation Counterfactuals

Hugh A. Graham

belian.earth · ML4EO 2026

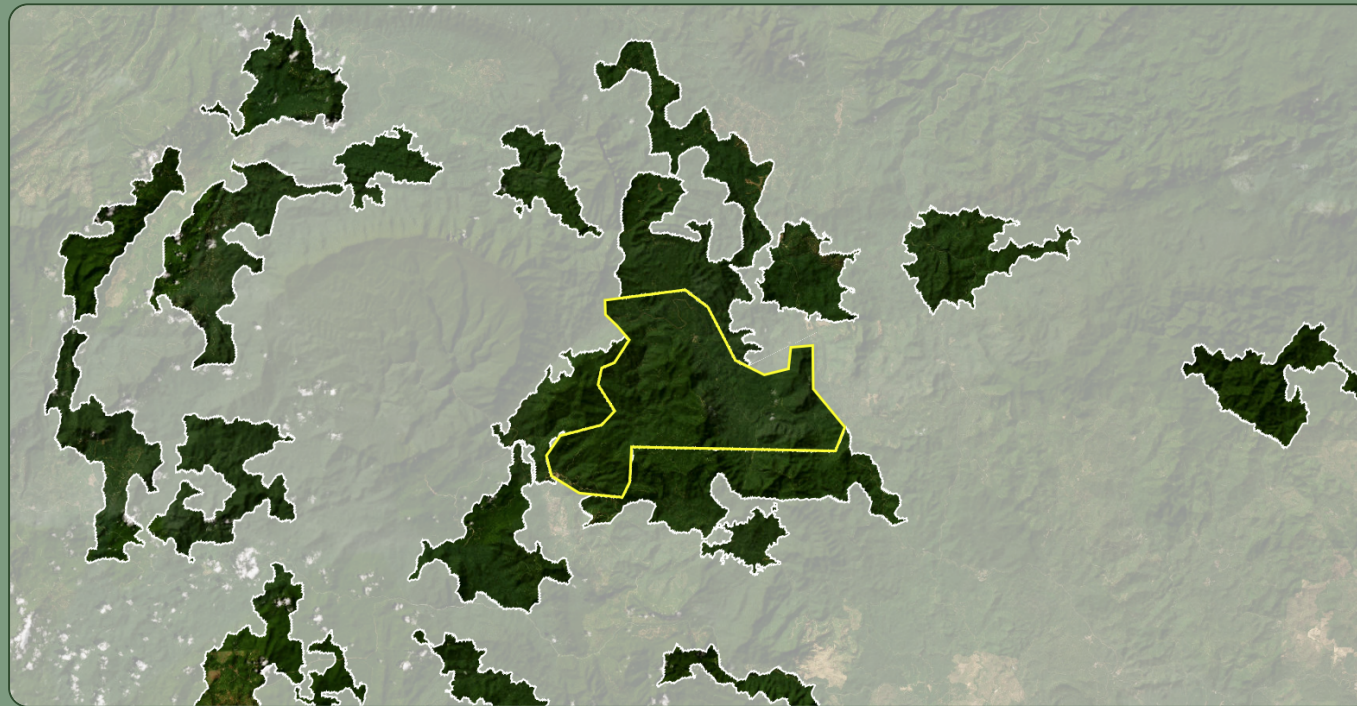
Christopher D. Philipson



# Find me a region like this

Reference-region selection is the weak point in every counterfactual.

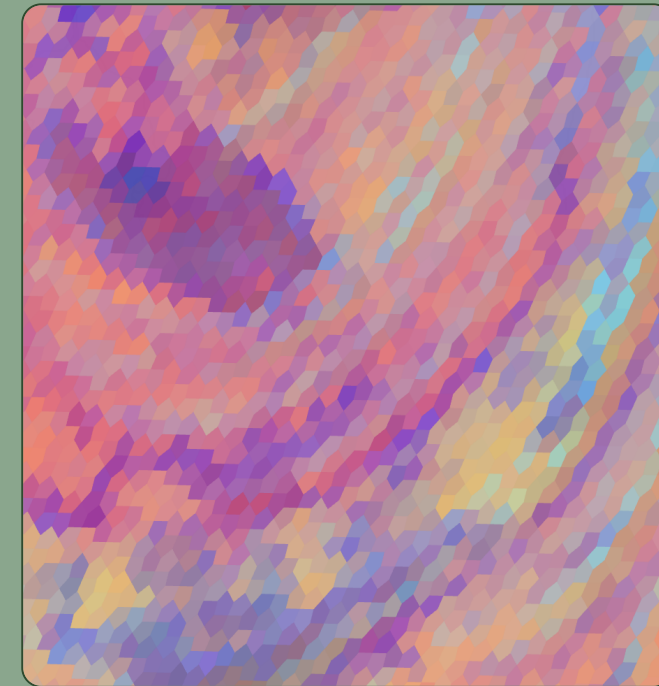
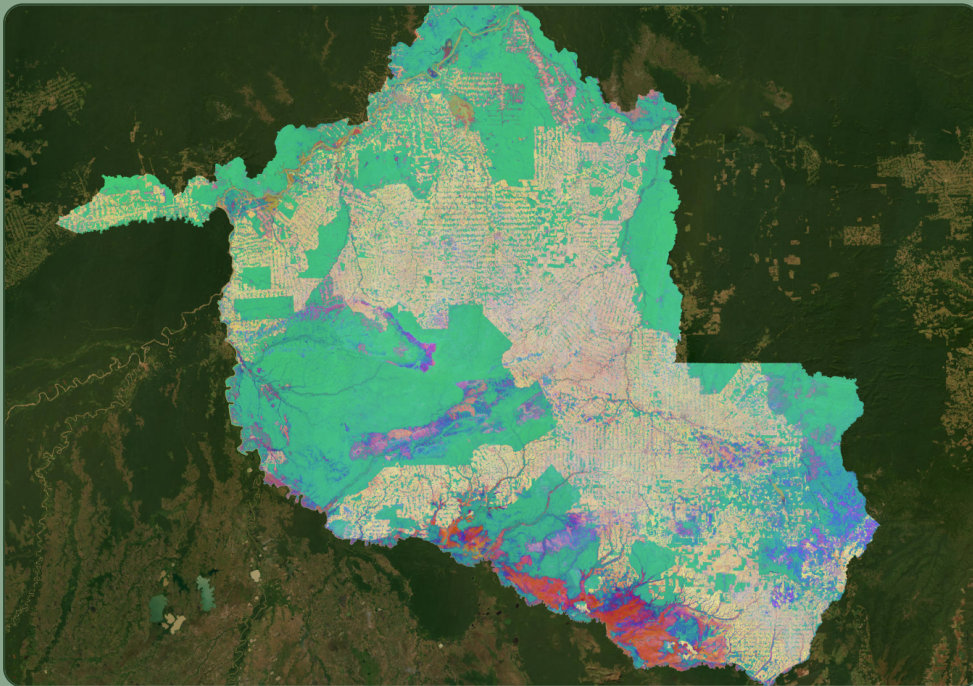
- ◆ Foundation model embeddings make *“find me a region like this”* directly computable
- ◆ This removes analyst discretion from donor selection: **explicit, reproducible, auditable**



# Embeddings on an equal-area grid

Every cell is an equivalent unit of comparison, stored as a searchable vector.

- ◆ **A5 pentagonal DGGS:** equal-area cells, comparable regardless of latitude
- ◆ **Parquet + LanceDB:** cell-keyed embeddings in columnar storage with vector similarity search and relational queries; a data format more capable than a served raster

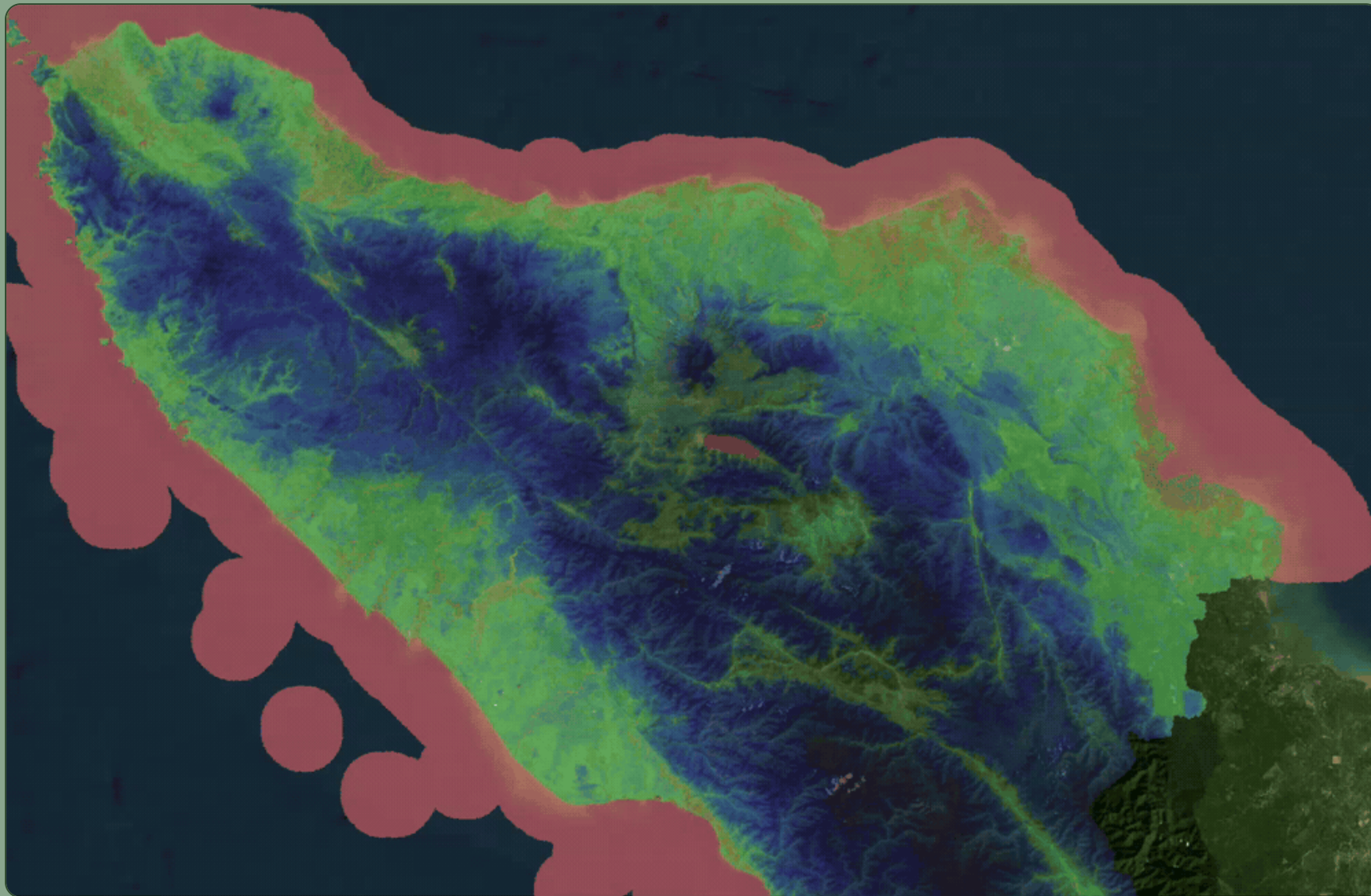


Embeddings: AlphaEarth Foundations, Brown et al. (2025), [arXiv:2507.22291](https://arxiv.org/abs/2507.22291)

A5 geospatial index: [a5geo.org](https://a5geo.org)



# Instant similarity search



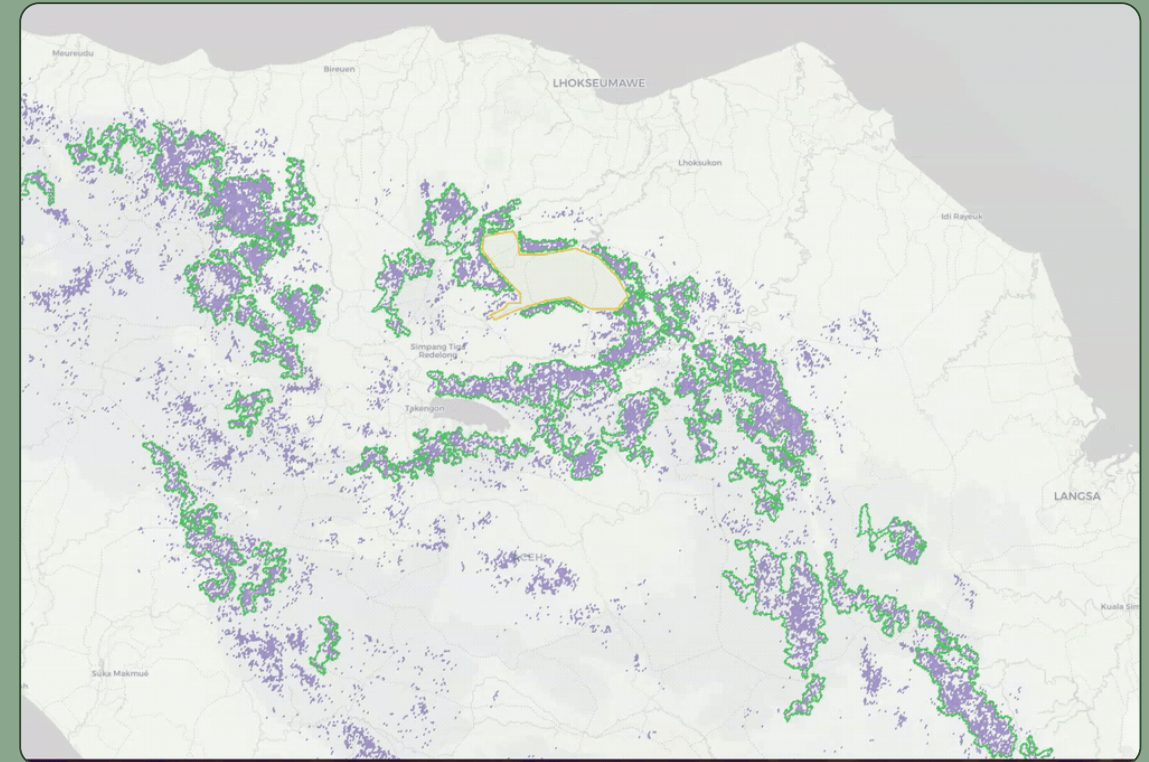
*Aceh, North Sumatra*



# Clustering into coherent regions

HDBSCAN turns scattered cell matches into contiguous reference regions.

- ◆ Restores the **contiguity and spatial autocorrelation** that cell/patch-wise matching discards
- ◆ Working at the cell level sidesteps the **Modifiable Areal Unit Problem (MAUP)**: arbitrary polygons average away the structure that makes cells comparable, inviting spurious matches



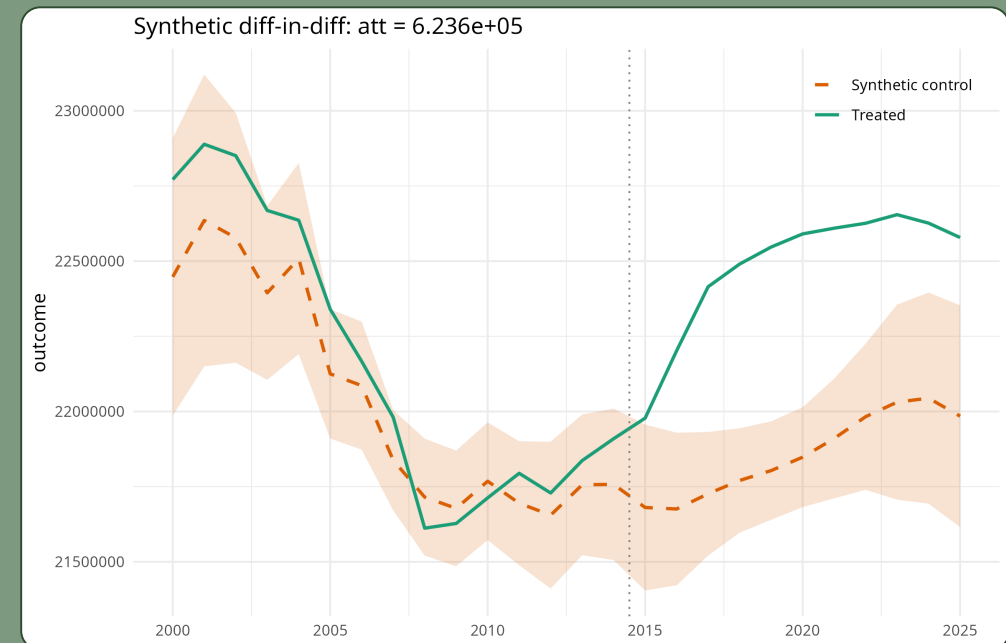
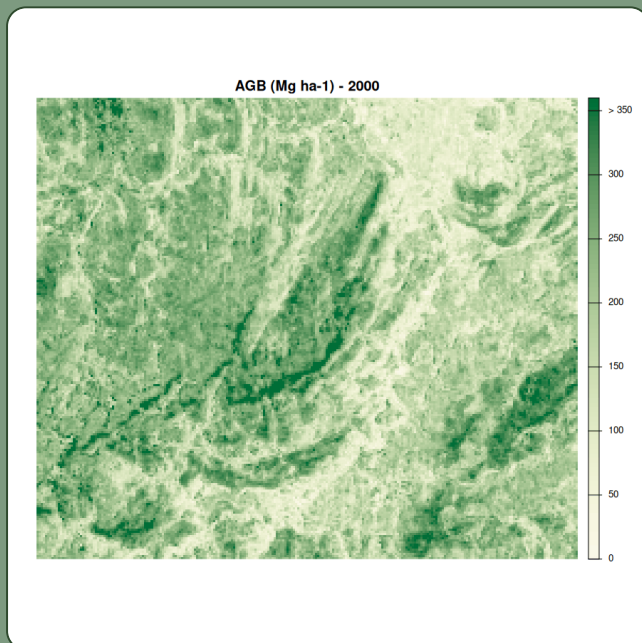
AI expands what is *searchable*; geospatial governs what is *comparable*.



# What works

Where the project and its donors share the same deforestation process, we can isolate the intervention's causal impact.

- ◆ Structural, recurring drivers (e.g. selective harvest across similar concessions): donors trace the project's business-as-usual
- ◆ **Synthetic difference-in-differences** on embedding-matched donors: clear improvement over an arbitrary-polygon baseline



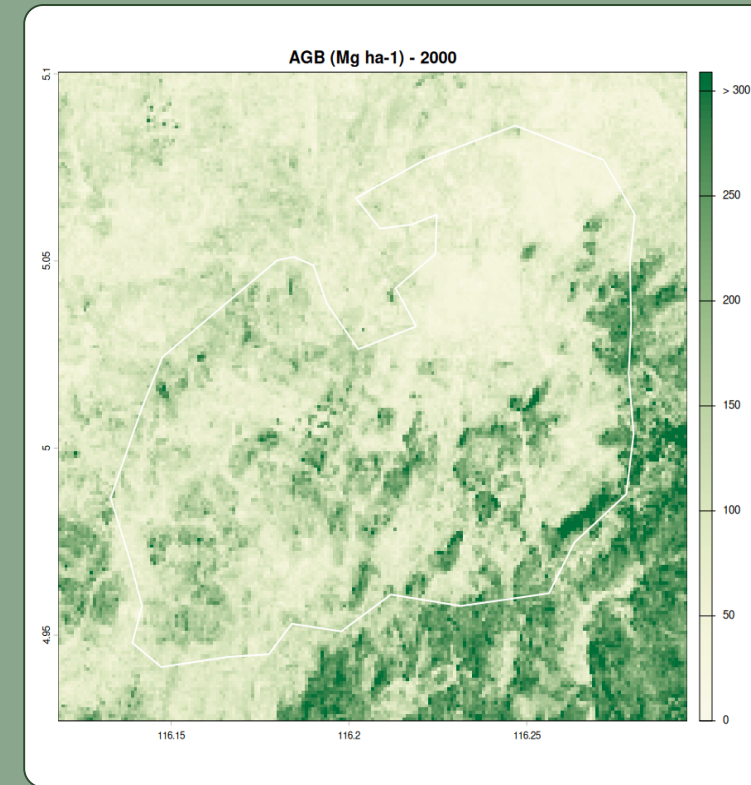
# The complication

We fixed *where* to compare, but we can't control volatility.

- ◆ We solve both problems in baseline construction: the statistical one (MAUP) and the analyst-discretion bias
- ◆ But a volatile, heavy-tailed event in the project need not coincide with one in the donor pool, and matching can't predict what the donors never sampled

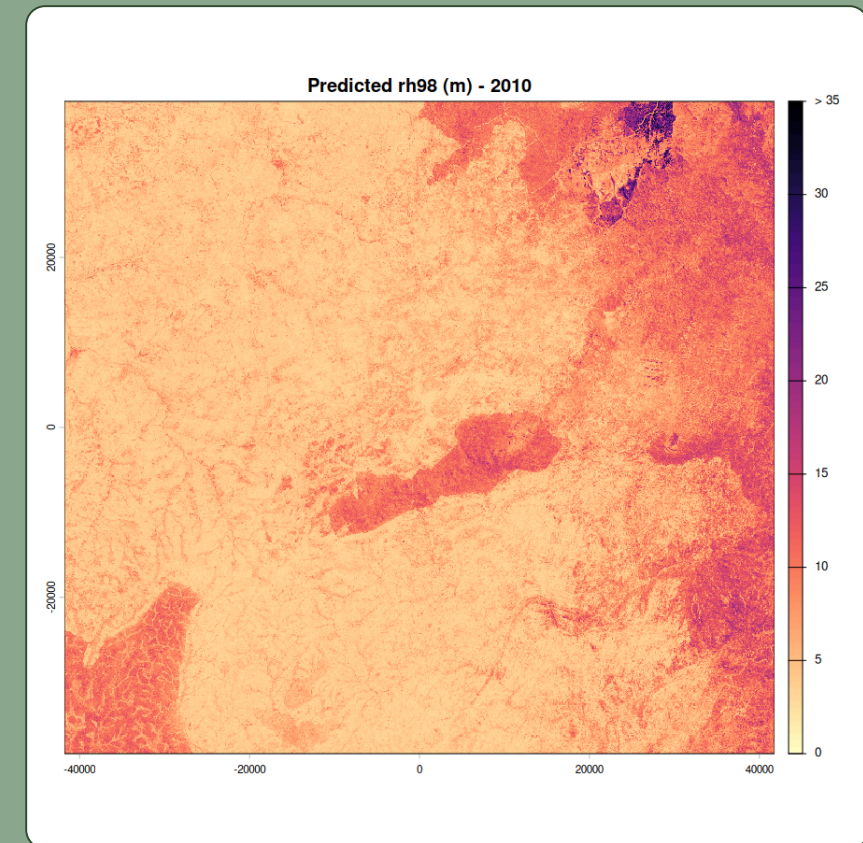
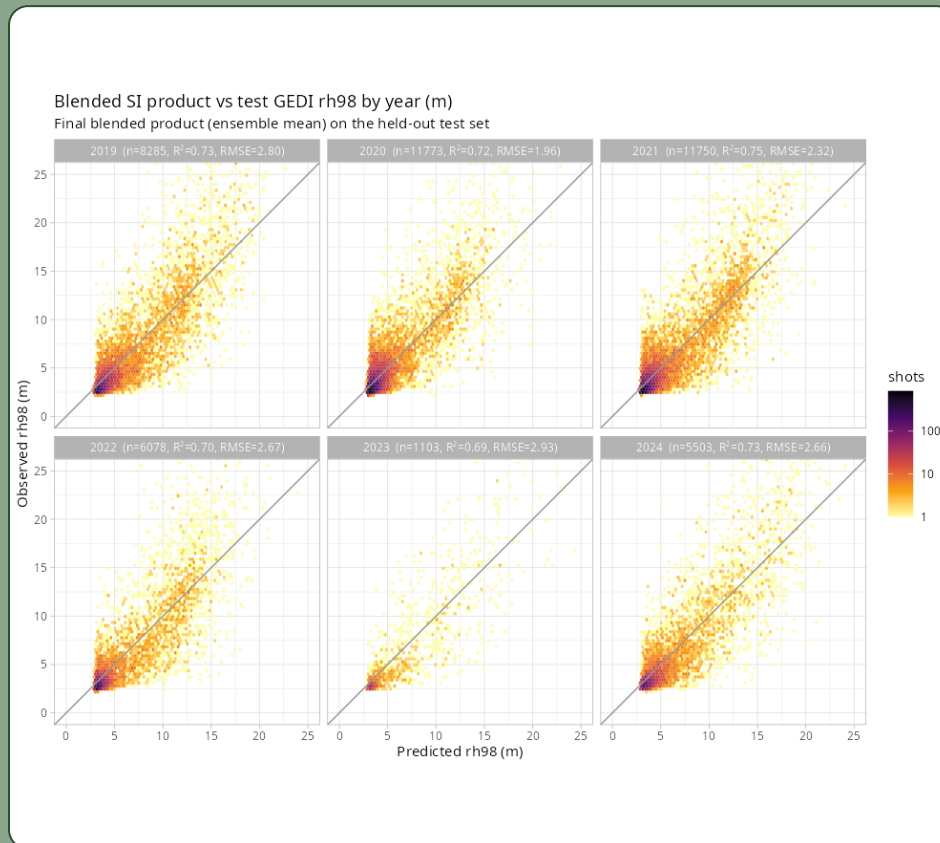
*Not a failure of matching: rare events are sparse in space and time. The information isn't in the donor sample.*

Lomax, Cunliffe, et al. (in review), *Synthetic controls underestimate deforestation in high-loss tropical forests.*



# Embeddings are amazing (but not a free lunch)

AlphaEarth and Embedded Seamless Data fusion for canopy height (RH98) modelling.



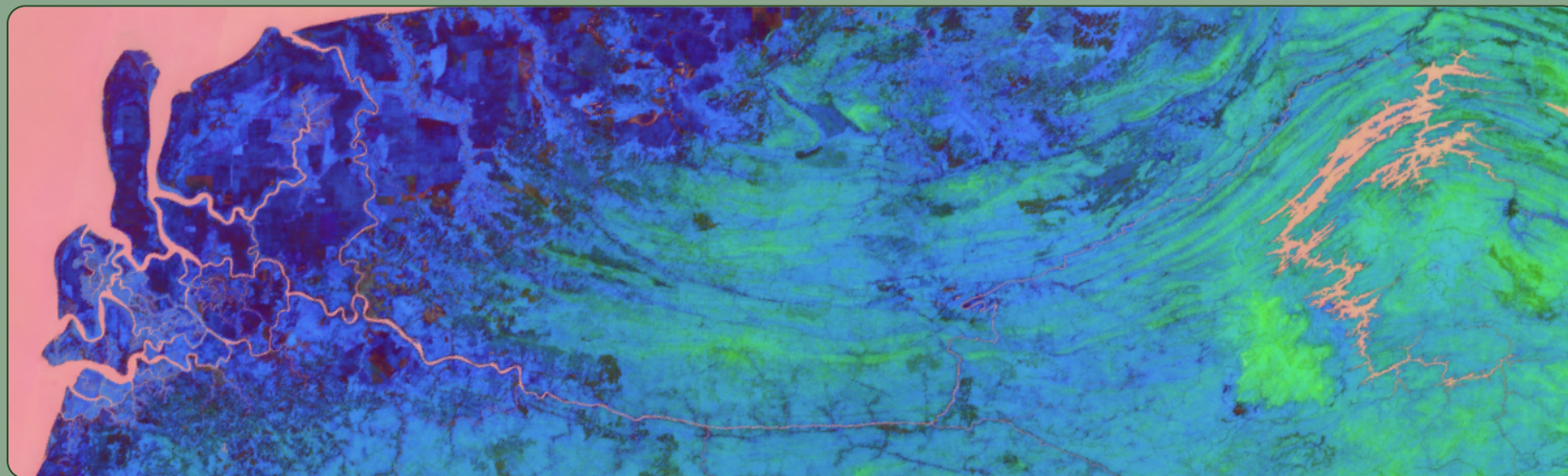
Embedding database: Chen et al. (2026), *Democratizing planetary-scale analysis: an ultra-lightweight Earth embedding database for accurate and flexible global land monitoring*, [arXiv:2601.11183](https://arxiv.org/abs/2601.11183)



# Where next

Honest intervals in the areas that matter most.

- ◆ Fuse **ex-ante** deforestation risk with **ex-post** SDiD in a regime-switching model
- ◆ Let intervals widen honestly where risk is high, rather than engineer false precision
- ◆ Validate against placebos: the bar is beating a wide null model *with real sensitivity*



# Thank you

Hugh Graham · Christopher Philipson

## Get in touch

 [belian.earth](https://belian.earth)

 [linkedin.com/company/belian-earth](https://linkedin.com/company/belian-earth)

 [github.com/belian-earth](https://github.com/belian-earth)

