Do Foundation Models Earn Their Labels?

Small-Data Biophysical Parameter Regression Under Geographic Shift

Sergei Kurashkin · Vadim Tynchenko · Vladimir Bukhtoyarov · Aleksei Borodulin

Bauman Moscow State Technical University, Moscow, Russia · kurashkin@bmstu.ru

CDEL: 2nd Workshop on Curated Data for Efficient Learning · ECCV 2026, Malmö

ECCV 2026 | MALMÖ | SEPT 8–12

The claim we test

Pre-training is offered as a substitute for annotation in label-scarce Earth observation. We test that premise at the budgets a curation decision turns on: how many labels a foundation model needs before it beats k-nearest neighbours on seven spectral indices, and whether the advantage survives a move to a new continent.

A label-efficiency curve is a statement about a protocol.

Corpus and grid

105,000
GEDI / Sentinel-2 patches

7
continents, 1,358 cells

100–10,000
labels per training draw

5
seeds per cell

9,975
evaluated units

6
systems compared

Coordinates are dropped at preparation time and their absence is verified automatically, so no model can memorise geography.

Six systems, one feature set

- **tabpfn3:** TabPFN v3, a tabular foundation model, on the seven indices.
- **knn, ridge, xgb, mlp:** the same seven indices, no pre-training and no learned representation.
- **ssl4eo:** a frozen SSL4EO-S12 ResNet-50, 2048 dimensions read out by a linear probe.

The indices are NDVI, NDMI, NDWI, NBR, red-edge NDVI, SAVI and EVI. Every tabular system reads the same rows, so a difference is attributable to the prior rather than to a richer feature space.

Protocol as a variable

105,000 patches · 7 continents
1,358 cells of 0.1°, balanced 15,000 per region

↓

loro: leave one continent out
the held-out fold is an entire continent, so the model predicts where it has seen no labels at all

inblk: in-region, blocked
whole 0.1° cells move together, so training and test patches are never adjacent

inrnd: in-region, random
retained only as the optimistic control

↓

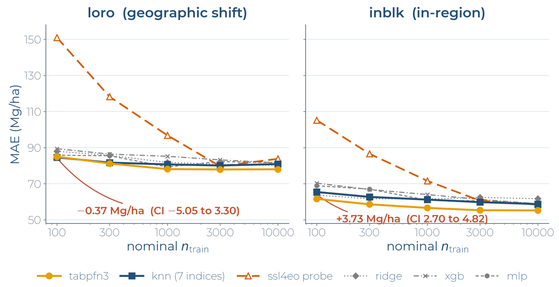
Paired bootstrap over 35 blocks + conformal coverage

Eight gate conditions

- **A1** a single clean pass with the resume mechanism disabled.
- **B1** every comparative claim backed by a second corpus, or a written waiver.
- **C1** the calibration metric present, and finite, in every unit.
- **D1** an optimism gap recorded for every system that tunes anything.
- **E1** the statistics files exist.
- **F1** one frozen feature set per regime.
- **G1** no coordinate feature anywhere in the reported grid.
- **H1** three TabPFN version tokens resolving to three distinct checkpoints.

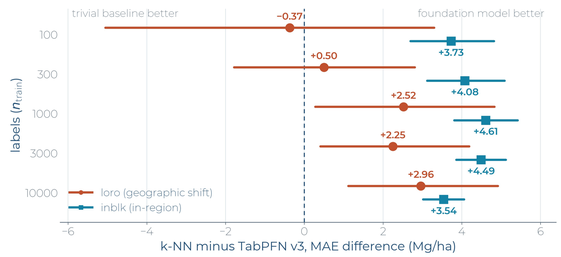
No number reaches a figure unless the gate passes. It is written before the final pass, so it cannot be tuned to let a result through. All eight pass on the frozen run of 9,975 units.

Label efficiency, two protocols



MAE against nominal training size, averaged over regions and seeds. The annotated values are the paired k-NN minus TabPFN v3 contrast at 100 labels.

The contrast, with intervals



Paired bootstrap across 35 blocks (7 regions × 5 seeds). Under shift the interval contains zero at 100 and 300 labels; in-region it never does.

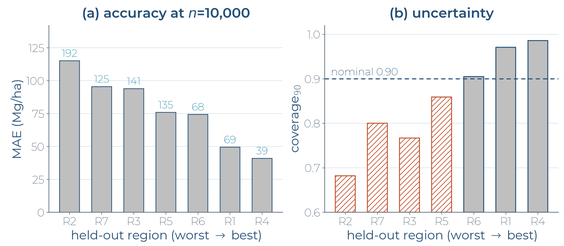
100 labels ≈ 725 labels

In-region, TabPFN v3 at 100 labels matches k-NN at roughly 725: a sevenfold saving in the unit a curation budget is denominated in. Under geographic shift at 100 labels the saving is absent.

An independent label pipeline

BioMassters, 15,000 crops from 166 Finnish chips labelled by airborne LiDAR rather than by GEDI, reproduces the in-region lead at 100 labels: +1.76 Mg/ha (95% CI 1.10 to 2.51). Its five blocks carry less weight than the 35 behind every AGBD contrast, and we report it as such.

Where and how it fails



tabpfn3 under loro, per held-out continent; the figure above each bar is that region's mean biomass. Hatched bars fall below nominal coverage: the interval fails in the same regions as the prediction. R1 Europe, R2 North Asia, R3 Australasia, R4 Africa, R5 South Asia, R6 South America, R7 North America.

−66.4 Mg/ha

The frozen SSL4EO-S12 probe at 100 labels, against the same seven indices. The deficit holds in-region too (−39.8 Mg/ha) and closes only near n≈3000.

What makes a region hard

Across the seven held-out continents, MAE tracks the region's mean biomass (Spearman rho = 0.86, permutation p = 0.024), whereas it does not track the region's label error (rho = 0.36, p = 0.44). With seven regions the coefficients are descriptive.

What a curation budget buys

- Pre-training buys labels inside the geography the corpus already covers, and stops at its edge.
- A budget aimed at a new region is better spent acquiring labels in it.
- Report the protocol with the curve: ours cross at n≈1000 under shift and never cross in-region.
- Put a cheap domain baseline in the comparison: it is what the budget is spent against.
- Measure conformal coverage rather than assume it: it falls where accuracy falls.

Limitations: one target variable, one label source for the cross-region claims, and a single frozen backbone, so the probe's deficit does not separate representation from readout.

Corpus preparation, grid, runner, frozen results and the gate: doi.org/10.5281/zenodo.21885451

code and data