Life of Research · Published 27 November 2026 · DOI 10.5281/zenodo.24573054

A Portable Encoder Is Not a Portable Rule: What Frozen Earth-Observation Embeddings Carry Across Regions, Why Recovery From a Few Target Labels Does Not Locate a Cross-Region Loss, and the Controls That Would Separate Readout, Shift and Place

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Global embedding products such as AlphaEarth Foundations and TESSERA are meant to be used frozen, with a small classifier fitted to local labels, and they are judged almost entirely by how well that local classifier does. A September 2026 wildfire study shows that the embeddings' advantage over Sentinel-2 index summaries inside a fire is mostly lost when the probe is fitted on fires in other regions and applied to a new one, that eight labelled blocks of the new fire restore it, and that the probe directions fitted in different fires are close to orthogonal; from this it concludes that the representation is not the cause. An August 2026 study of the same AlphaEarth product finds the reverse ordering: in Victoria, Australia, embedding models transferred to nearby regions far better than models on physical variables. In this analysis the author reviews and synthesises published results to show that the two studies, and the product papers behind them, use the word transfer for three different properties - an encoder that supports a good local rule anywhere, a rule that survives a move to a new region, and a rule that survives a move to a later year - and that the published evidence supports the first, contradicts an unconditional form of the second, and is a single pair of fires for the third. Recovery from a few target labels establishes only that the target is linearly separable in the embedding; that is predicted equally by regional differences in what burns, by differences in what the control pixels are, and by a high-dimensional probe fitting the place a particular fire happened to cover. The wildfire study's own tables add a fact its conclusion leaves out: with identical labels, the index summaries lose 0.055 AUC on transfer and the embeddings 0.148 to 0.176, so the extra loss is attached to the descriptor, not to the task. The paper consolidates the measured values in a table and a figure, sets out four accounts of a lost margin with the evidence that bears on each, and specifies an attribution procedure - coordinate and placebo-footprint controls, composition-matched backgrounds, a pooled shared-rule ceiling and a physical-anchoring check - that would say which account holds.

Keywords: geospatial foundation models; Earth-observation embeddings; spatial transferability; linear probing; distribution shift; wildfire susceptibility; evaluation methodology; domain adaptation

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A new mode of using satellite data has become routine in two years. Instead of training a model on imagery for each task, a practitioner downloads a precomputed vector for every 10 m pixel on Earth and fits a small classifier to whatever labels are at hand. The products that make this possible - AlphaEarth Foundations from Google DeepMind (Brown et al., 2025) and TESSERA from Cambridge (Feng et al., 2025) - descend from a decade of representation learning for remote sensing, from Tile2Vec's spatial version of the distributional hypothesis (Jean et al., 2018) through masked-autoencoder and contrastive pretraining on multispectral and multi-temporal imagery (Cong et al., 2022; Wang et al., 2022; Jakubik et al., 2023; Tseng et al., 2023; Szwarcman et al., 2024; Tseng et al., 2025). The idea that one encoding can serve many tasks if each user fits only a linear map to local ground truth was argued directly by Rolf et al. (2021), and a recent statement of what an ideal Earth foundation model should provide sets out eleven features such a model would need and how they should be evaluated (Zhu et al., 2026).

How these products should be evaluated is not settled. GEO-Bench asks for results from fine-tuned pretrained backbones, with frozen-backbone results optional, on randomly subsampled chips (Lacoste et al., 2023); PANGAEA finds that a supervised UNet often matches or beats the foundation models, and on its one cross-region split every model loses between 45% and 67% of its score (Marsocci et al., 2024); an audit of 152 papers finds 47 cross-paper disagreements of five points or more within matched model-benchmark-protocol groups and concludes that nobody can currently rank these models (Corley et al., 2026). The literature on spatially structured data has argued for years that random or weakly blocked validation overstates how well a map predicts at unvisited places (Roberts et al., 2017; Ploton et al., 2020; Kattenborn et al., 2022; Karasiak et al., 2022), and that assessing maps built from sparse reference data is itself an open problem (Meyer and Pebesma, 2022). Remote sensing has studied the failure of classifiers moved between images and regions as a domain adaptation problem for at least as long (Tuia et al., 2016), and a recent position paper argues that the

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

spatial and geographic character of satellite data calls for methods and evaluations designed around it (Rolf et al., 2024).

Into this setting a September 2026 study by Stark, Sopasakis and Hall puts a clean and well-controlled result (Stark et al., 2026). Using Copernicus burned-area maps of four 2023 fires in Greece and Spain, with TESSERA and AlphaEarth embeddings from the year before each fire, they find that a logistic probe inside a fire separates burned from surrounding pixels 0.05 to 0.13 AUC better than Sentinel-2 index summaries, under 21 block-to-fold allocations, buffers, a block bootstrap and gradient boosting. Applied to a fire in another region, the embedding probes lose 0.15 to 0.18 AUC on average and the index probes 0.06, so the three end level. Eight labelled blocks of the new fire put the embeddings ahead again; adding 59,000 pixels from the other fires to those blocks makes the embedding probe worse; and the probe direction fitted in one fire has a cosine of 0.03 to 0.41 with the direction fitted on the others. Their reading is stated in six words: "The representation is not the cause."

An August 2026 study of AlphaEarth for wildfire susceptibility in Victoria, Australia, points the other way (Zhuang et al., 2026). There, models on embeddings transferred to two nearby regions with a relative ROC-AUC change of about +4% and -2%, against a mean drop of about 25% for models on harmonised physical variables. Agricultural benchmarks of the same product report limited spatial transferability compared with conventional remote-sensing features (Ma et al., 2025) and cross-region accuracy of about 0.3 for irrigated-cropland mapping against cross-year accuracy above 0.87 (Yang et al., 2026).

In this analysis the author reviews and synthesises published results to show that this apparent conflict dissolves once three properties that share the word transfer are separated, and that the wildfire study's attribution of its own loss is stronger than its evidence. The paper argues five things.

1. Encoder portability, rule portability and temporal portability are distinct claims (Section 2). The product papers establish the first; the wildfire and agricultural studies refute an unconditional form of the second; the evidence on the third is thin and mixed.
2. Recovery from a handful of target labels establishes that the target is linearly separable in the embedding, which excludes only the strongest form of representation failure. It is predicted equally by three other accounts of a lost margin - regional differences in what burns, differences in what the control pixels are, and a probe fitting the particular place a fire covered (Sections 4 and 5).
3. Near-orthogonal probe directions are weak evidence of anything in 64 or 128 correlated dimensions; the functional evidence the wildfire study also reports - held-out accuracy peaking at a few principal components while within-fire accuracy climbs to full rank - is more informative, and it points to local fitting of high-variance directions (Section 6).
4. The loss is attached to the descriptor. With identical labels and design, the index summaries lose about a third as much as the embeddings, so whatever changed between regions changed the embedding-label relation more than the index-label relation. Under domain adaptation theory the quality of the best shared rule is a property of the representation, which makes "the representation is not the cause" the wrong summary of what was shown (Section 7).
5. The Victoria result does not contradict the Mediterranean one once the label, the scale, the split and the comparison descriptor are lined up, and in both studies the descriptor with the richer local fit lost more on transfer, with a caveat about model families (Section 8).

The strongest objection to this paper is that the wildfire authors already flag most of these caveats themselves. They do flag several - control composition, six fires, a single temporal pair - and their Appendix B says outright that "the transfer numbers reflect how alike the burned-control contrasts of two fires are, not how much a representation knows about fire." The contribution here is to show that this sentence and the abstract's attribution cannot both be the summary, to place the result against the representation-versus-head literature in machine learning where the same inference has been made and qualified (Kirichenko et al., 2022; Rosenfeld et al., 2022; Izmailov et al., 2022), and to specify controls that would decide it. An earlier paper by the author asked what random, spatial and site-disjoint splits each estimate and how many held-out sites a general claim needs (Mahendrakar, 2026). That question is taken as answered here; this paper asks a different one - given a measured cross-region loss under a sound split, where the loss lives.

2. Three Properties Called Transfer

Write phi for a frozen encoder that maps a pixel's satellite record to a vector, h for a head fitted on labelled pixels, and f_r for the labelling rule in region r - here, which pixels burned in a given fire. Three properties are in play.

Encoder portability (T1). For any region r, there exists a head h_r, fitted to modest local labels, such that h_r composed with phi predicts f_r well. This is what the product papers claim and test. AlphaEarth's evaluation fits k-nearest-neighbour or linear heads to at most 300

points per class within each of 15 evaluation datasets, with sample points thinned to at least 1.28 km apart to limit spatial autocorrelation, and reports an average error reduction of about 23.9% against the next-best featurisation in that setting (Brown et al., 2025). TESSERA reports that retraining the encoder on region-specific data offers negligible gains, so that one global encoder serves diverse geographies, and its section on geographic generalisation compares foundation models on crop datasets from Togo and Ethiopia (Feng et al., 2025). Rolf et al. (2021) state the paradigm in its purest form: the encoding is computed once, and each user fits a linear regression to their own ground truth. National-scale applications follow the same pattern, fitting light classifiers to labels inside the mapped country (Elbarz et al., 2026; Koukos et al., 2026). None of these claims says that a head fitted in one place works in another.

Rule portability (T2). A head h_A fitted in region A, composed with phi, predicts f_B in region B nearly as well as a head fitted in B. This is what the wildfire study, the Victoria study and the agricultural benchmarks test, and what an operational user without local labels needs.

Temporal portability (T3). A head fitted in region A on one year's event predicts a later event in A. This is what a forecasting user with a mapped history needs, and what the wildfire study's Attica pair tests.

The distinction matters because evidence for T1 is routinely read as evidence for T2. A benchmark that fits a head and tests it on held-out points from the same dataset measures T1 with some spatial separation; PANGAEA's cross-region split on FiveBillionPixels is one of the few standard tests of T2, and on it every frozen foundation model lost between 45% and 67% of its mIoU, while the end-to-end UNet lost 50% (Marsocci et al., 2024). The PANGAEA authors note that five minor classes present in training are absent from the cross-region test, so part of that loss is a change in the label distribution, not only in the inputs. An agricultural benchmark that assigns training, validation and test to separate regions within each of four US states finds the fine-tuned models collapse on minority crops, with SpectralGPT's crop mIoU ranging from 25.95 in North Carolina to 7.45 in California (Shang et al., 2026); that study reports only the held-out-region numbers, so the size of the loss relative to an in-region fit cannot be read from it, and its authors note they were still auditing one model's evaluation.

The classical theory of learning across domains gives T2 a precise shape. Ben-David et al. (2010) bound the target error of a rule by its source error, a divergence between the source and target input distributions in the representation, and a term lambda: the combined error, on both domains, of the single best rule in the hypothesis class. When lambda is large, no rule over phi does well in both places, and no amount of source data or divergence reduction rescues transfer. Zhao et al. (2019) sharpened this: aligning the input distributions is not sufficient when the class-conditional distributions differ between domains, and when label marginals differ there is an information-theoretic floor on joint error for any method built on invariant representations. The point for what follows is that lambda is a joint property of the representation and the two labelling rules. A finding that rules differ across regions is a finding about lambda, and lambda is not independent of phi.

3. What the Wildfire Study Reports

The design is careful, and the numbers below are its own (Stark et al., 2026). The four benchmark fires are Rhodes and Evros in Greece, Caceres in mainland Spain and Tenerife in the Canary Islands, all from 2023, with burned areas from 9,595 to 93,939 ha. For each, 5,000 burned pixels are sampled inside the mapped polygons after a 30 m inward buffer, and 15,000 controls from the activation area outside a 500 m exclusion buffer; after excluding TESSERA pixels with no embedding, 79,056 pixels enter every analysis. Four descriptors are compared: ESA WorldCover 2021 class, 15 Sentinel-2 index summaries (NDVI, NDMI and NBR median, interquartile range, 10th and 90th percentiles and observation count), the 128-dimensional TESSERA embedding, and the 64-dimensional AlphaEarth embedding, the last three from 2022. The probe is L2-regularised logistic regression with balanced class weights; cross-validation holds out complete 5 km blocks, with 39 to 90 blocks per fire, over 21 block-to-fold allocations.

Within fires, measured against all 15 index summaries including the observation counts, TESSERA leads by $+0.061$ to $+0.132$ AUC and AlphaEarth by $+0.053$ to $+0.105$, positive in nearly every allocation. Removing training pixels within 1 km of the test fold lowers every descriptor by a similar amount; gradient boosting keeps a TESSERA lead of $+0.042$ to $+0.119$. The study also reports two practical traps it found and avoided: scikit-learn's grouped stratified splitter produced only three distinct splits from 21 seeds on Rhodes, and TESSERA's all-zero no-data vectors occurred only among controls, a label leak worth up to 0.004 AUC in this sample.

Across fires, using a probe fitted on the other three, the mean held-out AUC is 0.690 for TESSERA, 0.707 for AlphaEarth and 0.711 for the index summaries without counts, against within-fire means of 0.865, 0.854 and 0.765 (Table 2 of the study). The per-fire losses are 0.128 to 0.230 for TESSERA, 0.105 to 0.177 for AlphaEarth and 0.024 to 0.089 for the indices. The sign of the held-out embedding-minus-index difference varies by fire, and its block-bootstrap interval excludes zero only on Evros and Caceres.

Three diagnostics follow. In the adaptation experiment, with eight labelled target blocks (11 to 26% of the target's training blocks), a TESSERA probe trained on

those blocks alone beats an index probe trained the same way by +0.12 on Rhodes and +0.04 on each other fire; at sixteen blocks the margins are +0.13, +0.06, +0.09 and +0.05. Adding the other three fires to the target blocks makes the TESSERA probe worse from eight blocks on, by up to 0.10 AUC at sixteen. In the direction analysis, with features standardised jointly across fires, the within-fire TESSERA direction has a cosine of 0.12 to 0.22 with the other-fires direction (AlphaEarth 0.03 to 0.41), and 91 to 99% of the within-fire coefficient norm lies in the fire-specific remainder; the remainder alone reproduces within-fire AUC and the shared direction alone gives roughly the held-out numbers. For the index summaries the cosines are 0.54, 0.24, -0.25 and 0.71, and the shared direction carries most of the within-fire AUC. Finally, a TESSERA probe on the top two to sixteen principal components, or regularised at $C = 10^{-4}$, beats the default full-rank probe held out on every fire; choosing C by leaving out a source fire raises the held-out TESSERA mean from 0.69 to 0.73, level with the indices, while within-fire AUC keeps rising to all 128 dimensions (the study's Section 4.3 puts the effect of this tuning at no more than 0.02; see Section 7).

The temporal test uses two fires in Attica, Parnitha (2023) and Varnavas (2024), twelve kilometres apart. A TESSERA probe fitted on Parnitha reaches 0.772 AUC on Varnavas's folds, 0.040 below a probe fitted inside Varnavas; single fires from other regions give 0.628 to 0.721. The reverse direction loses 0.29 AUC for TESSERA. For AlphaEarth, Parnitha gives 0.740 and Tenerife 0.758; for the index summaries, Parnitha gives 0.649 and Evros 0.790.

Table 1 collects these values with the comparison results from other studies on which the argument draws.

Figure 1 draws the two quantities from the wildfire study that carry most of the argument: the within-fire and held-out AUC of each descriptor in each fire, and the agreement between fitted probe directions.

4. Four Accounts of a Lost Margin

Suppose a descriptor beats another inside a region and loses that advantage when its rule is moved. Four accounts predict this, and they are not exclusive.

A1. The representation lacks the transferable information. The encoder does not encode what makes pixels burn in a way that is available in the target. This is the account the wildfire study sets out to exclude.

A2. What burns differs between regions. The probability that a pixel with a given pre-fire land surface burns depends on things the land surface does not contain: the weather of the fire days, where the fire started, how it was fought. Fire ecology says the relation between vegetation and burning itself varies regionally; on an aridity gradient across Iberia, fire activity tracked the sensitivity of fire to flammable conditions, which depended on fuel structure, more than the frequency of those conditions (Pausas and Paula, 2012). Cross-regional fire-occurrence models in the Alps and the Mediterranean transferred well only between environmentally similar regions (Bekar et al., 2020), and explainable deep models of fire danger in the eastern Mediterranean found that input contributions varied across fire events in a way that hints at different mechanisms (Kondylatos et al., 2022). The forecasting datasets built for this region carry daily weather and human-activity drivers precisely because land surface alone does not determine burning (Prapas et al., 2021; Kondylatos et al., 2023; see also Huot et al., 2021). Put in the terms of the causal-learning literature, which ties what can transfer to the causal direction between inputs and label (Schoelkopf et al., 2012), the pre-fire surface is one cause among several; when the unmodelled causes change between regions, the conditional of label on surface changes, and no encoding of the surface alone removes that.

A3. The control pixels differ. The task is not burned against unburned land in general; it is burned against what was available within an activation area outside a 500 m buffer. The control class is a different mixture in each fire. The study itself observes that Tenerife's burned area is almost entirely pine forest while its controls are a mixture of shrub, bare lava, built-up land and coast, so that WorldCover class alone reaches 0.81 AUC there. Species distribution modelling has documented this dependence for decades: AUC under a presence-background design varies with the geographic extent from which the background is drawn (Lobo et al., 2008), and how far from known presences pseudo-absences are drawn is itself a modelling decision (VanDerWal et al., 2009). In the language of shift, the class ratio is fixed at 1:3 by design, so this is not label shift in the sense of Lipton et al. (2018); it is a change in the distribution of features within the control class, a conditional shift of the kind that defeats representation alignment (Zhao et al., 2019).

A4. The probe fits the place. A single fire perimeter is one realisation of an event: the polygon covers the land where this fire went, which is spatially contiguous and tens of kilometres across. A descriptor that resolves fine landscape differences - this valley's slope and aspect, this ridge's stand structure, this hillside's cultivation history - lets a probe learn which part of the activation area the perimeter covered, including through features that have nothing to do with flammability. Five-kilometre blocks with 39 to 90 blocks per fire leave most test blocks surrounded by training blocks carrying the same label, so a rule that interpolates landscape similarity across a few kilometres scores well within the fire. Structured data offer "ample opportunity for overfitting with non-causal predictors", as Roberts et al. (2017) put it; in central African forest biomass, a non-spatial validation credited a model with more than half the variance while spatial

Table 1. Measured values the argument rests on, one row per published result, with the source named on each row. AUC is ROC AUC; "rel." marks relative changes as the source defines them; AE = AlphaEarth, S2 = index summaries.

Source	Setting	Descriptors	Evaluation	Reported value
Stark et al. 2026, Tab. 1	4 fires, GR/ES	TESSERA vs S2 (15, with counts)	within fire, 21 allocations	+0.061 to +0.132 AUC
Stark et al. 2026, Tab. 1	same	AlphaEarth vs S2 (15, with counts)	within fire	+0.053 to +0.105 AUC
Stark et al. 2026, Tab. 2	same	TESSERA / AE / S2	loss, within minus held-out	mean 0.176 / 0.148 / 0.055
Stark et al. 2026, Tab. 2	same	TESSERA / AE / S2	held-out mean AUC	0.690 / 0.707 / 0.711
Stark et al. 2026, Sec. 4.4	same	TESSERA vs S2, 8 target blocks	target-only probes	+0.12 (Rhodes), +0.04 others
Stark et al. 2026, Sec. 4.4	same	TESSERA, 16 blocks + 59k source px	vs target blocks alone	worse by up to 0.10 AUC
Stark et al. 2026, Tab. 6	same	TESSERA / AE / S2	cosine, fire vs other fires	0.12-0.22 / 0.03-0.41 / -0.25-0.71
Stark et al. 2026, Sec. 4.5	same	TESSERA, C chosen by LOFO	held-out mean AUC	0.69 to 0.73 (Sec. 4.3: change at most 0.02)
Stark et al. 2026, Tab. 3; Sec. 4.6	Attica 2023 to 2024	TESSERA	forward / reverse loss	0.040 / 0.29 AUC
Zhuang et al. 2026, Fig. 7	Victoria to Canberra, WSBM	AEF vs physical vars	rel. AUC change, 4 families	+4%, -2% vs about -25%
Zhuang et al. 2026, Tab. S6	Victoria, random 70/30	physical vs AEF, RF	within AUC (%)	95.67 vs 91.94
Yang et al. 2026	Guanzhong, Kansas	AEF, irrigated cropland	cross-year / cross-region OA	above 0.87 / about 0.3
Marsocci et al. 2024, Tab. 11	FiveBillionPixels	12 frozen GFMs; UNet	rel. mIoU change, regional gap	-45% to -67%; -50%
Shang et al. 2026, Tab. 5	4 US states	SpectralGPT (fine-tuned)	held-out crop mIoU	25.95 (NC), 7.45 (CA)
Brown et al. 2025	15 eval datasets	AEF vs next best	error reduction, within dataset	about 23.9%
Kumar et al. 2022	10 shift datasets	fine-tune vs linear probe	ID / OOD accuracy	+2% / -7% on average
Zhang et al. 2026	BR, NG, US	PDFM vs covariates	geo-structured CV, unexpl. var.	median -20.1%

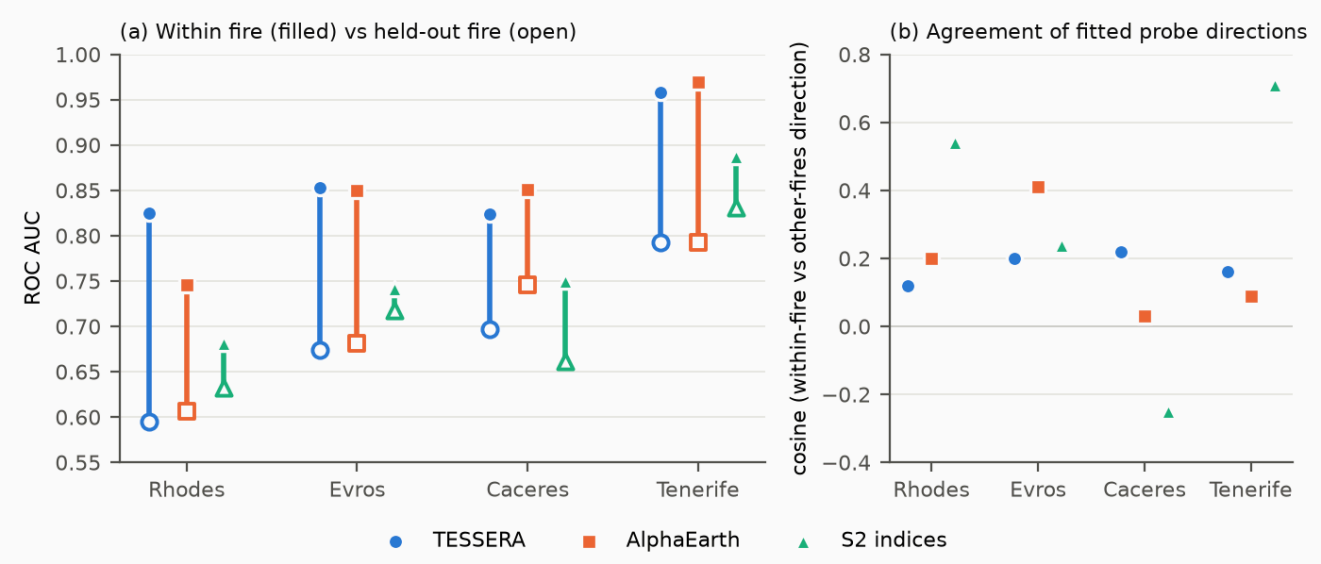

Figure 1. (a) Within-fire (filled) and held-out-fire (open) ROC AUC for TESSERA, AlphaEarth and the Sentinel-2 index summaries without observation counts, logistic probe, mean over 21 allocations; the vertical bar is the transfer loss. (b) Cosine between the probe direction fitted inside each fire and the direction fitted on the other three fires (allocation 0, mean of five folds). All values from Stark et al. (2026), Tables 2 and 6.

validation found quasi-null predictive power (Ploton et al., 2020); spatially autocorrelated training and validation samples inflate the assessed accuracy of convolutional networks (Kattenborn et al., 2022), and spatial dependence between training and test sets is a recognised pitfall of accuracy assessment in remote-sensing classification (Karasiak et al., 2022). A4 is not a claim that the within-fire result is leaked in the usual

sense. The 1 km buffer test rules out leakage at the scale of a kilometre. It is a claim that the label varies at a scale larger than the block, so that within-region validation, however blocked, rewards descriptors that can fit place.

The four accounts make different predictions, and the wildfire study's diagnostics touch them unequally. Table 2 summarises which published observation bears on which account; Section 10 states the controls that would separate them.

5. What Recovery From a Few Target Labels Establishes

The adaptation experiment is the centre of the study's attribution. Its authors state its logic plainly: poor transfer can arise because the representation lacks information for the target or because the rule learned elsewhere does not apply there, and recovery with limited target supervision indicates the former is false.

The same inference has a history in machine learning. Kirichenko et al. (2022) showed that classifiers relying on spurious backgrounds still encode core features, since retraining only the last layer on data where the spurious correlation is broken recovers much of the lost robustness. Rosenfeld et al. (2022) argued that standard training may already learn features sufficient for out-of-distribution generalisation, so that a small amount of target data and a retrained linear layer give excellent performance, and proposed a domain-specific adjustment into a canonical latent space before a shared linear predictor. Izmailov et al. (2022) used last-layer retraining as a measuring instrument for feature quality and found that quality depends heavily on architecture and pretraining choices. In each case the logic is the wildfire study's logic: if a linear readout fitted on target-distribution data succeeds, the information was in the features.

That inference is valid for what it claims, and what it claims is narrow. It shows that the target labels are linearly separable in the representation, given target-distribution training data. It excludes A1 in its strong form - that the encoder lacks the information in the target region - and nothing else. A2 predicts recovery: if each region has its own rule linking surface to burning, and the encoder resolves the surface, a local readout will find the local rule. A3 predicts recovery: if the controls differ, a local readout learns the local contrast. A4 predicts recovery most directly of all: a probe fitted on eight blocks of the target learns which landscapes in the target lie under the perimeter, and is scored on other blocks of the same perimeter, which is the within-region measurement whose limits the study itself demonstrates.

That last point deserves care, because it is where the recovery test is weakest. The adaptation probe is scored on held-out folds of the target fire. It is therefore a within-region number, with the spatial separation of 5 km blocks, and the study's headline is that within-region numbers do not predict accuracy in a new region. A recovery that is itself measured within the region cannot discriminate between a representation that carries portable fire information read out by the wrong rule and a representation that offers enough local degrees of freedom to fit any spatially coherent contrast.

The probing literature in natural language processing faced this problem and named the missing control. Hewitt and Liang (2019) observed that a probe's accuracy reflects both what the representation encodes and what the probe can memorise, and proposed control tasks - arbitrary labels with the same structure as the real task - so that a probe's selectivity, real accuracy minus control accuracy, measures the representation rather than the probe. Voita and Titov (2020) found that probe accuracy did not substantially favour pretrained over random representations and could be similar for genuine and random labels unless probe size or data were constrained, and replaced accuracy with description length. Belinkov (2021) reviews these and other shortcomings of probing, including the gap between what a probe can decode and what a model uses. The wildfire adaptation test has no control task. The natural one is a placebo footprint: a polygon of the same shape and size placed on unburned land in the same activation area, with controls sampled by the same rule. If the embeddings separate placebo footprints from their surroundings with a margin over the indices comparable to the real one, the real margin is mostly a statement about how well the embeddings fit place.

The second half of the adaptation result - that adding 59,000 pixels from other fires makes the embedding probe worse from eight target blocks on - is informative but does not discriminate either. When two regions require different rules, a pooled fit compromises between them and loses to a local fit once the local data suffice; A2, A3 and A4 all predict that. The result does say something about lambda: within this hypothesis class, a single linear rule over the embedding that serves the source fires and the target is worse than the target-only rule, which is a direct sign that the shared-rule term of Ben-David et al. (2010) is large for linear heads on these embeddings. Section 7 returns to why that is not a fact about the labels alone.

A different strand qualifies the inference from the other side. Fine-tuning can distort pretrained features and do worse out of distribution than a linear probe when features are good and shift is large, with accuracy about 7% lower out of distribution and 2% higher in distribution on average across ten shift datasets (Kumar et al., 2022), and which layers are worth adapting depends on the kind of shift: for input-level corruptions, early layers, not the head (Lee et al., 2022). With a frozen encoder, only the head can move. A recovery achieved by moving the head shows that a head exists; it does not show that the shift lives at the head rather than in the input distribution that the encoder maps.

Table 2. What each published observation from the wildfire study implies for each account of the lost margin. "Consistent" means the account predicts the observation; "against" means the account predicts its absence. The classification is the author's.

Observation (Stark et al. 2026)	A1 no info	A2 what burns	A3 controls	A4 place
Within margin robust to buffers, boosting	neutral	consistent	consistent	consistent
8 target blocks restore margin	against	consistent	consistent	consistent
Pooling source fires hurts from k = 8	neutral	consistent	consistent	consistent
Near-orthogonal directions	neutral	consistent	consistent	consistent
Held-out peaks at few PCs; within climbs	neutral	weakly	weakly	consistent
Indices lose about a third as much	consistent	needs extra	needs extra	consistent
Tenerife easiest target, worst source	neutral	neutral	consistent	neutral
Attica forward 0.772, reverse much lower	neutral	mixed	mixed	mixed

6. What the Probe Directions Establish

The direction analysis supports the study's conclusion about the embeddings' local advantage: "It is a different rule in each fire." Two features of the evidence limit what can be drawn from the cosines themselves, and a third piece of evidence the study reports is more telling.

First, a weight vector is not a function. In a space of 64 or 128 correlated, standardised features, many weight vectors produce nearly the same scores; this is the underspecification that D'Amour et al. (2020) identify as a general source of predictors that are equivalent on held-out data in the training domain and behave differently in deployment. Two probes can therefore be far apart in weight space and close in what they predict, or the reverse. The study's functional checks - scoring the held-out fold with the shared direction, the remainder and the full direction - are the right complement, and they show that the remainder carries the within-fire AUC while the shared direction alone reproduces roughly the held-out AUC.

Second, raw cosines are not comparable across descriptors of different dimension. For directions drawn at random and isotropically in d dimensions, the expected absolute cosine is close to the square root of 2/(pi d). That gives about 0.07 for d = 128, 0.10 for d = 64 and 0.23 for d = 12; these are the author's arithmetic, offered as an order-of-magnitude reference, and the isotropy assumption is not met by correlated features. On that reference the TESSERA cosines of 0.12 to 0.22 are about 1.7 to 3 times the isotropic baseline for their dimension, while the index cosines of 0.24 and -0.25 on Evros and Caceres are about at the baseline for 12 dimensions. The reading that the index summaries share a rule across fires and the embeddings do not is right on Rhodes and Tenerife, where the index cosines are 0.54 and 0.71, and is not supported by the cosines on Evros and Caceres. Figure 1(b) shows the spread.

The more informative evidence is the dimensionality result. Held out, TESSERA probes on two to sixteen principal components, or with strong regularisation, beat the full-rank default on every fire, and leave-one-fire-out selection of the regularisation raised the held-out mean from 0.69 to 0.73 (though the study's Section 4.3 puts the tuning effect at no more than 0.02; see Section 7); within the fire, AUC keeps rising to full rank. The study draws the right first conclusion: the part of the embedding signal that carries across fires is small and low-dimensional. The second conclusion is that the high-dimensional part - the part that wins within the fire - is the part that does not travel. That is the signature A4 predicts: a coarse, shared contrast (greener, wetter, more wooded land burns more, in roughly the way the index summaries also capture) plus fine, high-variance directions that fit the local perimeter. A2 can produce the same signature if the regional rules differ in fine detail, which is why this evidence is consistent with A4 rather than decisive for it. It sits alongside an interpretability finding that 2 to 12 of AlphaEarth's 64 dimensions recover 98% of baseline land-cover classification performance depending on class (Benavides-Martinez et al., 2026). The group-robustness literature does not offer a parallel for the regularisation gain: there, strong L2 regularisation raised worst-group accuracy only when combined with an objective aimed at the worst group (group DRO), and under ERM alone worst-group accuracy stayed poor (Sagawa et al., 2019), whereas the wildfire probe is an ordinary ERM logistic regression.

The two products also differ in a way that bears on A4 and that the wildfire study does not exploit. TESSERA is pixel-wise: its pretraining uses a global shuffling regulariser described as decorrelating spatial neighbourhoods, so a pixel's vector is built from that pixel's own time series (Feng et al., 2025). AlphaEarth is an embedding field: its encoder interleaves a spatial self-attention operator, a time-axial attention operator and a 3x3 convolutional precision operator, so each pixel's vector carries information from its surroundings (Brown et al., 2025). If A4 worked mainly through neighbourhood context - a probe learning that pixels inside large contiguous stands sit under the perimeter - AlphaEarth should fit place more easily and lose more on transfer. It does not: TESSERA's mean loss (0.176) is larger than AlphaEarth's (0.148). That observation counts against the neighbourhood-context form of A4 and leaves its other form intact, in which a pixel's own multi-sensor annual trajectory is distinctive enough to separate one valley's land from the next. The Victoria study reports a related effect at 1 km: the gain from

convolutional patch models over cell-wise models was larger for embeddings than for physical variables (Zhuang et al., 2026), so spatial context supplied by the head helped there. Neither observation isolates the mechanism; together they show that the place account, if it holds, need not run through spatial context.

The concatenation result belongs here as well. Concatenating TESSERA and AlphaEarth gives the best within-fire AUC on every fire, by +0.006 to +0.047 over the better single embedding, and held out it is no better than either alone except on Tenerife. More local degrees of freedom bought more local fit and no portability. A benchmark of five geospatial models' frozen embeddings finds them neither redundant nor disjoint (Ouaknine and Rolnick, 2026), which explains why concatenation helps locally; the wildfire result shows that complementary local information is not portable information.

7. The Asymmetry the Attribution Leaves Out

The study compares four descriptors under one design. The labels, the sampling, the blocks, the probe and the source fires are identical. What differs is the descriptor, and the descriptors lose very different amounts when moved: a mean of 0.055 AUC for the index summaries, 0.148 for AlphaEarth and 0.176 for TESSERA, or roughly 2.7 and 3.2 times the index loss (author's arithmetic from the study's Table 2).

Consider what A2 and A3 alone would predict. If what burns differs between regions, or the controls differ, those differences live in the world and in the sampling, and they would degrade any descriptor that captures the relevant surface variables. The index summaries capture some of them - vegetation greenness, moisture, burn-sensitive reflectance - and they do lose: 0.089 on Caceres, 0.056 on Tenerife. But the embeddings lose much more on three of four fires under the same change of world and sampling. A2 and A3 can accommodate that only with an extra premise: that the embeddings' additional within-fire information is precisely the part of the surface whose relation to burning (or to the local control mixture) changes between regions, while the index summaries see only the part whose relation is stable. That premise may be true. It is a statement about which information the representation carries and how that information is arranged, which is to say about the representation.

A4 needs no extra premise. A low-dimensional descriptor of 12 index summaries cannot fit place at the resolution a 128-dimensional embedding can, so it has less local margin to gain and less to lose. Under A4, the asymmetry is the expected outcome.

Either way, the phrase "the representation is not the cause" is the wrong summary. What the study shows is that the information needed to separate burned from control pixels in each fire is present in the embeddings. What it also shows, in its Table 2 and its Section 4.5 and Figure 5, is that the best single linear rule over the embeddings across fires is no better than the default linear rule over 12 index summaries (the 15 without the three observation counts) - 0.73 against 0.71 held out once the embedding probe is strongly regularised (the index probe at its default regularisation). The study elsewhere puts the effect of leave-one-fire-out tuning at no more than 0.02 (its Section 4.3 and Appendix C), so the tuned TESSERA held-out mean lies between about 0.71 and 0.73: level with the indices, not ahead of them. In the terms of Section 2, the representation supports T1 well and does not improve T2 over a hand-crafted baseline in this task. Under Ben-David et al. (2010), that is a statement about lambda, and lambda is a joint property of the representation and the labelling rules. Whether the labelling rules differ "really", in the world, or only as seen through this representation, is exactly the open question; the study's evidence does not decide it, and its abstract decides it.

A more precise summary of what was shown would read: the embeddings contain fire-specific separating directions that a local probe can find from a few labelled blocks, and contain a shared direction about as good as the shared direction in the index summaries. That sentence is supported by every table in the study. It is also less quotable, and it does not tell a user without local labels anything encouraging.

8. The Other Side: Where Embedding Rules Did Travel

The Victoria study is the strongest published evidence that an embedding-based rule can travel better than a conventional one (Zhuang et al., 2026). Models were trained in Victoria on satellite-derived fire occurrence over 2017 to 2025, aggregated to a 1 km grid, with a 1:1 balance of fire and non-fire cells, using tree ensembles, multilayer perceptrons, TabPFN and convolutional networks on either 64-dimensional AlphaEarth embeddings or physical variables drawn from a pool of 35 harmonised predictors, of which 22 were retained after collinearity screening. Transferred without modification to eight sites across Australia, embedding models lost little at the two nearest temperate sites - relative ROC-AUC change of about +4% at Canberra and -2% at Western Sydney-Blue Mountains - while physical-variable models dropped by about 25% on average there. Losses grew for both representations with climatic distance, largest at the arid Alice Springs site, and recall fell faster than specificity.

Lined up against the Mediterranean study, five differences stand out, and each bears on the accounts of Section 4.

First, the label. Victoria's positives are cells with any fire occurrence over nine years across a whole state; the Mediterranean positives are pixels inside one fire's

perimeter, contrasted with that fire's surroundings. A multi-year, state-wide occurrence label averages over many ignitions and weather sequences and is closer to susceptibility as fire ecology means it; a single perimeter is one draw. A4 predicts that a one-event label rewards fitting place far more than a many-event label does.

Second, the scale. Embeddings averaged to 1 km lose the fine within-cell detail that a 10 m probe can exploit. If A4 is right, aggregation removes much of what a local fit could latch onto.

Third, the split. The within-Victoria evaluation is a stratified random 70/30 split repeated ten times, and the transfer change is computed relative to that within-region score. Random splitting of spatially structured data inflates the within score (Roberts et al., 2017; Ploton et al., 2020), for both representations, so the relative transfer change partly measures how much each representation benefited from the random split.

Fourth, the comparison descriptor. Victoria's physical variables include climate summaries - temperature, precipitation, wind, fire-weather indices - whose ranges shift between regions. A model fitted on Victoria's climate gradient and applied where those variables take values outside its training range is extrapolating; the area-of-applicability framework of Meyer and Pebesma (2021) was built to flag exactly that, and the Mediterranean study's 15 index summaries contain no climate variables to extrapolate on.

Fifth, the within-region ordering. In Victoria the tree ensembles favoured the physical variables within region - for random forest, a test ROC-AUC of 95.67% for physical variables against 91.94% for embeddings (Zhuang et al., 2026, Table S6) - while the multilayer perceptron, convolutional and TabPFN models sat close to parity. Averaged across the four families, the physical-variable models then lost more on transfer. In the Mediterranean study, the embeddings had the larger within-fire margin and lost more. Read together, and with the caveat that Victoria's transfer numbers are averaged over model families whose within-region orderings differ, the two studies are consistent with a single pattern: the descriptor that fit the source region more closely lost more when moved, whether it was the embedding or the hand-crafted set. That reading suggests that the question "do embeddings transfer better than conventional features?" has no general answer, because the answer tracks which descriptor overfits the source region's particular contrast in a given design.

A matched comparison would settle whether the two studies disagree at all. It would take the Victoria labels and region pair, replace the random 70/30 split with blocked validation inside Victoria, add a single-event version of the label (one season's perimeters against their surroundings), run the cell-wise probes at 10 m and at 1 km, and swap the 22 retained physical variables (screened from a 35-variable pool) for a set without climate variables, such as the Mediterranean study's index summaries. Each change moves one design factor from one study's value to the other's. If the embedding advantage on transfer survives all of them, the Victoria result is a property of the embeddings in temperate southeastern Australia; if it disappears with the single-event label or the 10 m scale, the two studies agree and differ only in design. The comparison needs no new satellite products, only the two studies' labels and descriptors under one protocol.

The other published transfer results fit the three-property distinction. For irrigated cropland, AlphaEarth gave stable cross-year accuracy above 0.87 within region and cross-region accuracy of about 0.3, which the authors attribute to differences in irrigation regimes, crop phenology and management - an A2-type explanation (Yang et al., 2026). The agricultural benchmark of Ma et al. (2025) found AlphaEarth competitive with purpose-built models when trained on local data but with limited spatial transferability compared to remote-sensing features. For local climate zones across five Swiss cities, a location-aware attention U-Net on embeddings transferred across cities well, with TESSERA ahead of AlphaEarth and of Sentinel-1/2 composites, while year-to-year transfer was named an open challenge (Ko Ko and Atzberger, 2026). For population estimation, foundation-model embeddings reduced unexplained variance by a median of 20.1% under geographically structured validation, unevenly across places and coupled to the spatial scale of training (Zhang et al., 2026). For Danish tree species, hand-built spectral-temporal features with a multilayer perceptron were best overall, with TESSERA within 1.1 points for pure stands and ahead when fewer than about a quarter of training plots were available (Koukos et al., 2026) - a T1 result about label efficiency. Location encoders such as SatCLIP claim to improve geographic generalisation by encoding visual similarity between distant places (Klemmer et al., 2023), which is a T2 hope rather than a demonstrated property of pixel embeddings used with a linear head.

The general machine-learning literature on distribution shift says the same thing at a higher level. Across many models and several natural shifts, including satellite classification in FMoW-WILDS, out-of-distribution accuracy tracks in-distribution accuracy closely (Miller et al., 2021), but the correlation weakens on some shifts, and on real-world datasets it is sometimes inverse, so that selecting by in-distribution accuracy misses the best out-of-distribution models (Teney et al., 2022). WILDS includes geographic shifts precisely because standard training shows a persistent in-distribution to out-of-distribution gap there (Koh et al., 2020), and evaluations of self-supervised transfer find that a standard benchmark ranking predicts many-shot recognition well and few-shot and dense tasks less well (Ericsson et al., 2020), while conclusions drawn under the linear evaluation protocol shift with architecture choices it

holds fixed (Kolesnikov et al., 2019). The wildfire study is a small instance of a broken line: three descriptors, four fires, an in-region ordering that does not survive the move.

9. Within-Region Forecasting Rests on One Pair

The study's most practically attractive claim is that a region with one mapped fire can forecast susceptibility for later fires there. The evidence is one pair of fires, and the study says so in its limitations. Three features of the pair limit what it can carry.

The asymmetry is large. Forward, from Parnitha to Varnavas, TESSERA loses 0.040; in reverse it loses 0.29, and whole-fire AUC is 0.738 forward and 0.619 in reverse. A property of the embeddings that held within Attica should not depend this strongly on direction.

The local advantage is specific to TESSERA. For AlphaEarth, Tenerife (0.758) was a better source for Varnavas than Parnitha (0.740); for the index summaries, Evros (0.790) and Rhodes (0.757) both beat Parnitha (0.649). Pooling the four remote fires gave TESSERA 0.765, close to the 0.772 from Parnitha alone.

The accounts make conflicting predictions here, and the pair splits them. A pure A4 account - the probe fits only the place under one perimeter - predicts weak transfer to a different perimeter even 12 km away, and the forward TESSERA result is evidence against it in that direction; the reverse result is evidence for it. A2 predicts good transfer within a region of similar fuels and climate, which the forward result supports and the reverse result complicates.

Other evidence on temporal portability is mixed in the same way. Cross-year transfer of irrigated-cropland rules was stable within region (Yang et al., 2026); year-to-year transfer of local-climate-zone models was not (Ko Ko and Atzberger, 2026). Annual embeddings summarise a year's observations, and TESSERA's accuracy depends strongly on the observation window for phenology-driven classes - contracting from one year to one month cost 39% of segmentation accuracy on PASTIS-R but 5% on DynamicEarthNet (Guerrero-Viu et al., 2026) - so a rule fitted on one year's embeddings meets a slightly different embedding of the same land in the next. The honest summary is that within-region temporal transfer of a frozen-embedding rule has been observed once for wildfire, in one direction, for one of two embeddings.

10. An Attribution Procedure

The accounts in Section 4 can be separated with the data the wildfire study already holds and a few more fires. Algorithm 1 states the procedure; Table 3 states what each account predicts under each control. No values appear in Table 3: it is a set of predictions, not a record of measurements.

Step 1 restates the study's own recommendation that every evaluation report a held-out region, and adds that the loss be reported beside both numbers. Step 2 borrows a control the AlphaEarth evaluation already runs: coordinates alone (XY) and with elevation (XYZ) as featurisations, described there as testing whether location "Is All You Need" (Brown et al., 2025); the wildfire study did not include one. Step 3 is the control task of Hewitt and Liang (2019) translated into geography. Step 4 implements the matched or land-cover-stratified controls that the wildfire study names in its limitations, and connects to the presence-background literature on background extent (Lobo et al., 2008; VanDerWal et al., 2009). Step 5 estimates the lambda term of Ben-David et al. (2010) for linear heads directly. Step 6 extends a reconstruction analysis that the Victoria study ran in one region - many physical predictors recovered from AlphaEarth with median R^2 above 0.8 (Zhuang et al., 2026) - to a cross-region comparison. Step 7 follows the advice that geographic transfer be related to measured shift: a study of mobility models across 2,265 US counties found transfer heterogeneous and asymmetric, with feature-distribution and spatial-structure shift each explaining part of it (Zhou et al., 2026), and the nearest-neighbour distance matching of Mila et al. (2022) gives a way to make validation distances resemble prediction distances. Unlabelled methods that predict out-of-distribution accuracy from agreement between models could in principle stand in for labels in step 7 (Baek et al., 2022); that phenomenon was reported for neural network classifiers, and whether it holds for linear probes on frozen embeddings is not known.

Two remarks on the table. A1 predicts low placebo selectivity because a representation without fire information should not prefer real perimeters to placebo ones, and the adaptation result already argues against A1; it stays in the table to keep the comparison complete. A4's distinctive predictions are the coordinate result and near-zero placebo selectivity; if both come out against it, the within-fire margin is fire information, and the remaining question is A2 against A3, which composition matching decides.

What follows for benchmark design is modest and concrete. A benchmark that wants to say anything about rule portability needs held-out regions as a standard split, not a single optional experiment: GEO-Bench subsamples chips at random (Lacoste et al., 2023; Shang et al., 2026), PANGAEA tests regional transfer on one dataset of eleven (Marsocci et al., 2024), and the AlphaEarth suite thins points to 1.28 km within each dataset (Brown et al., 2025). The audit of Corley et al. (2026) asks for shared core evaluations, variance reporting and controls that separate data, architecture and algorithm; a held-out-region split with its within-region twin, and a coordinate featurisation as a standing control, fit naturally within those expectations.

Algorithm 1. Attributing a cross-region loss of a frozen-embedding advantage

```
Inputs: fires F (several per region), descriptors D (embeddings, indices,
        coordinates), probe family H, block size b
1. For each fire f and descriptor d: within-fire AUC W[f,d] on b-km blocks;
   held-out AUC T[f,d] from a probe fitted on fires outside f's region.
   Report W, T and the loss W - T; never W alone.
2. Coordinate control: repeat step 1 with d = (x, y) and d = (x, y, elevation).
   If W for coordinates approaches W for embeddings, within-fire scores
   largely measure place.
3. Placebo footprint: for each fire, place copies of its polygon on unburned
   land in the same activation area; resample controls by the same rule;
   compute the within-fire margin of each descriptor over the indices.
   Selectivity = real margin - mean placebo margin.
4. Composition matching: resample controls so that their land-cover
   composition matches the burned pixels within each fire; recompute W and T.
   The part of the loss removed is attributable to control composition (A3).
5. Shared-rule ceiling: fit one probe on all fires including labelled target
   blocks, with a fire indicator excluded; compare its target AUC with the
   target-only probe at the same number of target blocks. A small gap means
   a shared rule exists in phi (small lambda); a large gap means it does not.
6. Physical anchoring: per region, regress the index summaries (and terrain)
   on the embedding; test whether the fitted maps agree across regions.
   Stable maps with unstable fire rules place the change in the world (A2/A3);
   unstable maps place it in how phi arranges regions.
7. Repeat across many fires per region and many regions; model fire as the
   unit of inference, and relate loss to measured source-target shift.
Output: for each account, the share of the loss removed by its control.
```

Table 3. Predictions of the four accounts under Algorithm 1's controls. "Large" and "small" are relative to the observed loss; no measured values are implied.

Control (Algorithm 1 step)	A1 no info	A2 what burns	A3 controls	A4 place
Coordinates, within fire (2)	n/a	low W	low to mid W	W near embedding
Placebo selectivity (3)	low	high	high	near zero
Loss after composition matching (4)	unchanged	unchanged	much smaller	unchanged
Shared-rule gap, pooled vs local (5)	large	large	shrinks after (4)	large
Embedding-to-index maps across regions (6)	unstable	stable	stable	stable
Loss vs fires per region (7)	flat	falls with more regions	falls after (4)	falls with more events per place

Reporting the loss between the two numbers, per region, would let a reader see whether a new encoder's gain is a gain in local fit, in portability, or in both. Under the accounts of Section 4, a gain in local fit alone is exactly what a high-dimensional encoder would show whether or not it encodes anything portable, so benchmarks that report only within-region numbers reward the property least informative about deployment without labels.

The operational reading is equally concrete. A user with labelled events in the target region - a mapped fire, a field campaign - is in the setting where the embeddings' advantage is best supported: the wildfire study's adaptation experiment and every T1 evaluation agree that a light local head on the embedding does well with few labels, and the Danish tree-species results show the advantage is largest when labels are scarcest (Koukos et al., 2026). A user without target labels should not assume that a rule fitted elsewhere on an embedding beats one fitted on simple physical descriptors; on the published wildfire evidence it does not, and on the published agricultural evidence it often does worse (Ma et al., 2025), and cross-region accuracy for irrigated cropland on AlphaEarth was only about 0.3 (Yang et al., 2026). In the wildfire study, strong regularisation brought the embedding rule level with the index rule held out, not ahead of it.

11. The Strongest Case Against This Paper

Four objections deserve a direct answer.

The first is that A4 is a hypothesis this paper cannot test, and that the 1 km buffer and the block bootstrap in the wildfire study already address spatial leakage. The buffer addresses leakage at the scale of the buffer. A4 concerns a label that varies at the scale of a fire perimeter, tens of kilometres, with 5 km blocks; no buffer short of holding out the perimeter addresses it, and holding out the perimeter is the held-out-fire test. The paper does not claim A4 holds. It claims that the published evidence does not exclude it and that its exclusion is required before the within-fire margin is called fire information.

The second is that the asymmetry in Section 7 is mechanical: a weak descriptor cannot lose much because it has little to lose. That is correct and is part of the point. If the asymmetry is mechanical, it is mechanical because the embedding carries additional local information that does not travel, and the question of what that information is - fire-relevant regional variation (A2) or place (A4) - is the question the abstract of the wildfire study answers without testing.

The third is that the Victoria comparison is unfair, since its design differs in label, scale, split and baseline. That is why Section 8 lists the differences rather than treating the two studies as a contradiction. The claim that the descriptor with the closer source fit lost more in both studies is offered as a pattern consistent with A4, hedged by the averaging over model families in Victoria; it is not a measured regularity.

The fourth is that the three-property distinction is obvious to the field. It is easy to state, but the vocabulary in use blurs it. The AlphaEarth abstract describes the representation as "highly general" on the strength of heads fitted within each evaluation dataset (Brown et al., 2025), which is T1 evidence; the wildfire study opens by noting that a large body of work shows such embeddings serve as reusable features, which is also T1, before testing T2. A reader who takes "general" to mean that a fitted rule travels is reading T2 into T1 evidence. Making the distinction explicit costs little and keeps a T1 benchmark from being cited for a T2 deployment.

Readers should weight the arguments of this paper unequally. Section 7's asymmetry and Section 5's point that recovery is measured within region rest directly on the wildfire study's own tables and are the strongest claims here. Section 6's cosine reference rests on an isotropy assumption that correlated features violate, and should be read as an order of magnitude. Section 8's cross-study pattern rests on two studies with different designs.

12. What Is Not Known

It is not known whether frozen-embedding probes fitted on wildfire perimeters are selective in the sense of Section 5; no placebo-footprint or coordinate control has been reported for this task.

It is not known how much of the cross-region loss in the Mediterranean study is due to control composition; the study names this as a limitation and does not measure it.

It is not known whether the relation between pre-fire embeddings and burning differs between Greek, Spanish and Canarian fires in the world, or only as arranged in the embeddings; no study has compared embedding-to-physical-variable maps across these regions.

It is not known whether a frozen-embedding rule transfers within a region from one year's fire to the next in general. One pair of fires, in one direction, for one of two embeddings, is the evidence.

It is not known whether the Victoria result would survive spatially blocked within-region validation, a single-event label, or 10 m resolution, and it is not known whether physical variables without climate extrapolation would transfer as poorly as the 22-variable set did.

It is not known how transfer loss scales with the number of training regions. Going from one training fire to three raised held-out TESSERA AUC by about 0.01 and AlphaEarth by about 0.08 in the wildfire study, and four fires cannot say what forty would do; a cross-regional fire-occurrence model in the Alps and the Mediterranean outperformed regional models in transfer tests (Bekar et al., 2020).

13. Conclusion

A frozen Earth-observation embedding can be portable as an encoder without any rule fitted on it being portable as a model. The published evidence supports the first property for the current products and contradicts an unconditional form of the second; for wildfire it contains one observation of a third, temporal portability within a region. The Mediterranean wildfire study demonstrates the gap between the first two properties more cleanly than any previous work. Its evidence shows that the embeddings hold enough information to fit a good rule in each fire and that the best shared rule over them is no better than one over twelve index summaries. It does not show that the representation is not the cause, and its own tables show the extra loss attached to the representation. Recovery from a few target labels, near-orthogonal directions and a harmful pooled fit are each predicted by regional differences in what burns, by differences in the control mixture and by a probe fitting the place a fire covered. A coordinate control, a placebo footprint, composition-matched controls, a pooled shared-rule ceiling and a cross-region physical-anchoring check would separate them, with the fires and labels already in hand. Until they are run, an evaluation of a frozen embedding should report the within-region score, the held-out-region score and the loss between them, and should claim encoder portability from the first and rule portability only from the second.

References

Baek, C., Jiang, Y., Raghunathan, A., Kolter, Z. (2022). Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift. arXiv:2206.13089. doi:10.48550/arXiv.2206.13089.

Bekar, I., Tavsanoglu, C., Pezzatti, G. B., Vacik, H., Pausas, J. G., Bugmann, H., Petter, G. (2020). Cross-regional modelling of fire occurrence in the Alps and the Mediterranean Basin. International Journal of Wildland Fire, 29, 712-722. doi:10.1071/WF19158.

Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J. W. (2010). A theory of learning from different domains. Machine Learning, 79, 151-175. doi:10.1007/s10994-009-5152-4.

Benavides-Martinez, I. F., Guthrie, J., Arias, J. E., Garces-Gomez, Y. A., Guzman-Alvis, A. I., Portilla-Cabrera, C. V., Mondal, S., Allyn, A. J., et al. (2026). What on Earth is AlphaEarth? Hierarchical structure and functional interpretability for global land cover. arXiv:2603.16911. doi:10.48550/arXiv.2603.16911.

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., et al. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data.

arXiv:2507.22291. doi:10.48550/arXiv.2507.22291.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., et al. (2022). SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. arXiv:2207.08051. doi:10.48550/arXiv.2207.08051.

Corley, I., Lehmann, N., Robinson, C., Tseng, G., Fuller, A., Alemohammad, H., Shelhamer, E., Marcus, J., et al. (2026). No One Knows the State of the Art in Geospatial Foundation Models. arXiv:2605.12678. doi:10.48550/arXiv.2605.12678.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., et al. (2020). Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395. doi:10.48550/arXiv.2011.03395.

Elbarz, W., Bourriz, M., Simonneaux, V., Ait Abdelali, H., Bourzeix, F. (2026). National-Scale Crop Type Mapping in Morocco Using Google AlphaEarth Foundations Embeddings and the Fields of the World Field Boundary Map. arXiv:2611.12681. doi:10.48550/arXiv.2611.12681.

Ericsson, L., Gouk, H., Hospedales, T. M. (2020). How Well Do Self-Supervised Models Transfer? arXiv:2011.13377. doi:10.48550/arXiv.2011.13377.

Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M. C., Immitzer, M., et al. (2025). TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis. arXiv:2506.20380. doi:10.48550/arXiv.2506.20380.

Guerrero-Viu, J., Lopez-Cifuentes, A., Perez-Villar, I., Pacifici, F. (2026). Temporal Sensitivity Analysis of Tessera Embeddings. arXiv:2608.27175. doi:10.48550/arXiv.2608.27175.

Hewitt, J., Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., Chen, Y. (2021). Next Day Wildfire Spread: A Machine Learning Data Set to Predict Wildfire Spreading from Remote-Sensing Data. arXiv:2112.02447. doi:10.48550/arXiv.2112.02447.

Izmailov, P., Kirichenko, P., Gruver, N., Wilson, A. G. (2022). On Feature Learning in the Presence of Spurious Correlations. arXiv:2210.11369. doi:10.48550/arXiv.2210.11369.

Jakubik, J., Roy, S., Phillips, C. E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., et al. (2023). Foundation Models for Generalist Geospatial Artificial Intelligence. arXiv:2310.18660. doi:10.48550/arXiv.2310.18660.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., Ermon, S. (2018). Tile2Vec: Unsupervised representation learning for spatially distributed data. arXiv:1805.02855. doi:10.48550/arXiv.1805.02855.

Karasiak, N., Dejoux, J.-F., Monteil, C., Sheeren, D. (2022). Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. Machine Learning, 111, 2715-2740. doi:10.1007/s10994-021-05972-1.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing, 5, 100018. doi:10.1016/j.ophoto.2022.100018.

Kirichenko, P., Izmailov, P., Wilson, A. G. (2022). Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. arXiv:2204.02937. doi:10.48550/arXiv.2204.02937.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Russwurm, M. (2023). SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. arXiv:2311.17179. doi:10.48550/arXiv.2311.17179.

Ko Ko, H. Y., Atzberger, C. (2026). Exploring the potential of AlphaEarth and TESSERA embeddings for Fine-scale Local Climate Zone Mapping: A case study across five cities in Switzerland. arXiv:2606.20034. doi:10.48550/arXiv.2606.20034.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., et al. (2020). WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421. doi:10.48550/arXiv.2012.07421.

Kolesnikov, A., Zhai, X., Beyer, L. (2019). Revisiting Self-Supervised Visual Representation Learning. arXiv:1901.09005. doi:10.48550/arXiv.1901.09005.

Kondylatos, S., Prapas, I., Camps-Valls, G., Papoutsis, I. (2023). Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the Mediterranean. arXiv:2306.05144. doi:10.48550/arXiv.2306.05144.

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernandez-Torres, M.-A., Carvalhais, N. (2022). Wildfire Danger Prediction and Understanding With Deep Learning. Geophysical Research Letters, 49. doi:10.1029/2022GL099368.

Koukos, A., Kondylatos, S., Nord-Larsen, T., Nyborg, L., Tottrup, C., Grogan, K. (2026). Tree species mapping in Denmark: A comparison of spectral-temporal features with geospatial foundation model embeddings. arXiv:2609.03480. doi:10.48550/arXiv.2609.03480.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P. (2022). Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. arXiv:2202.10054. doi:10.48550/arXiv.2202.10054.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E. D., Kerner, H., Lutjens, B., Irvin, J. A., Dao, D., et al. (2023). GEO-Bench: Toward Foundation Models for Earth Monitoring. arXiv:2306.03831. doi:10.48550/arXiv.2306.03831.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., Finn, C. (2022). Surgical Fine-Tuning Improves Adaptation to Distribution Shifts. arXiv:2210.11466. doi:10.48550/arXiv.2210.11466.

Lipton, Z. C., Wang, Y., Smola, A. (2018). Detecting and Correcting for Label Shift with Black Box Predictors. arXiv:1802.03916. doi:10.48550/arXiv.1802.03916.

Lobo, J. M., Jimenez-Valverde, A., Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17, 145-151. doi:10.1111/j.1466-8238.2007.00358.x.

Ma, Y., Shen, Y., Swatantran, A., Lobell, D. B. (2025). Harvesting AlphaEarth: Benchmarking the Geospatial Foundation Model for Agricultural Downstream Tasks. arXiv:2601.00857. doi:10.48550/arXiv.2601.00857.

Mahendrakar, P. (2026). A Held-Out Site Is a Sample of One: What Random, Spatial and Site-Disjoint Splits Each Estimate, Why Leakage That Differs Between Configurations Reorders a Comparison, and How Many Sites a General Claim Needs. Zenodo. doi:10.5281/zenodo.23824501.

Marsocci, V., Jia, Y., Le Bellier, G., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., et al. (2024). PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models. arXiv:2412.04204. doi:10.48550/arXiv.2412.04204.

Meyer, H., Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12, 1620-1633. doi:10.1111/2041-210X.13650.

Meyer, H., Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Communications, 13. doi:10.1038/s41467-022-29838-9.

Mila, C., Mateu, J., Pebesma, E., Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. Methods in Ecology and Evolution, 13, 1304-1316. doi:10.1111/2041-210X.13851.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., et al. (2021). Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. arXiv:2107.04649. doi:10.48550/arXiv.2107.04649.

Ouaknine, A., Rolnick, D. (2026). Towards Consistent Geospatial Representations: Aligning One Encoder to Many Foundation Models. arXiv:2610.28313. doi:10.48550/arXiv.2610.28313.

Pausas, J. G., Paula, S. (2012). Fuel shapes the fire-climate relationship: evidence from Mediterranean ecosystems. Global Ecology and Biogeography, 21, 1074-1082. doi:10.1111/j.1466-8238.2012.00769.x.

Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11. doi:10.1038/s41467-020-18321-y.

Prapas, I., Kondylatos, S., Papoutsis, I., Camps-Valls, G., Ronco, M., Fernandez-Torres, M.-A., Piles Guillem, M., Carvalhais, N. (2021). Deep Learning Methods for Daily Wildfire Danger Forecasting. arXiv:2111.02736. doi:10.48550/arXiv.2111.02736.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40, 913-929. doi:10.1111/ecog.02881.

Rolf, E., Klemmer, K., Robinson, C., Kerner, H. (2024). Mission Critical – Satellite Data is a Distinct Modality in Machine Learning. arXiv:2402.01444. doi:10.48550/arXiv.2402.01444.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. Nature Communications, 12. doi:10.1038/s41467-021-24638-z.

Rosenfeld, E., Ravikumar, P., Risteski, A. (2022). Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization. arXiv:2202.06856. doi:10.48550/arXiv.2202.06856.

Sagawa, S., Koh, P. W., Hashimoto, T. B., Liang, P. (2019). Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. arXiv:1911.08731. doi:10.48550/arXiv.1911.08731.

Schoelkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J. (2012). On Causal and Anticausal Learning. arXiv:1206.6471. doi:10.48550/arXiv.1206.6471.

Shang, Z., Das, S., Eldawy, A. (2026). Benchmarking Geospatial Foundation Models for Agriculture Applications. arXiv:2606.29664. doi:10.48550/arXiv.2606.29664.

Stark, P., Sopasakis, A., Hall, O. (2026). When local gains fail to transfer: Frozen Earth-observation embeddings across wildfires. arXiv:2609.34602. doi:10.48550/arXiv.2609.34602.

Szwarcman, D., Roy, S., Fraccaro, P., Gislason, T. E., Blumenstiel, B., Ghosal, R., de Oliveira, P. H., de Sousa Almeida, J. L., et al. (2024). Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. arXiv:2412.02732. doi:10.48550/arXiv.2412.02732.

Teney, D., Lin, Y., Oh, S. J., Abbasnejad, E. (2022). ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets. arXiv:2209.00613. doi:10.48550/arXiv.2209.00613.

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H. (2023). Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv:2304.14065. doi:10.48550/arXiv.2304.14065.

Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J. R., Shelhamer, E., et al. (2025). Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. arXiv:2502.09356. doi:10.48550/arXiv.2502.09356.

Tuia, D., Persello, C., Bruzzone, L. (2016). Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. IEEE Geoscience and Remote Sensing Magazine, 4, 41-57. doi:10.1109/MGRS.2016.2548504.

VanDerWal, J., Shoo, L. P., Graham, C., Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling, 220, 589-594. doi:10.1016/j.ecolmodel.2008.11.010.

Voita, E., Titov, I. (2020). Information-Theoretic Probing with Minimum Description Length. arXiv:2003.12298. doi:10.48550/arXiv.2003.12298.

Wang, Y., Ait Ali Braham, N., Xiong, Z., Liu, C., Albrecht, C. M., Zhu, X. X. (2022). SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation. arXiv:2211.07044. doi:10.48550/arXiv.2211.07044.

Yang, L., Gao, Y., Zhao, X., Liang, N., Ma, R., Xi, S., Zhang, X., Wang, R. (2026). Evaluating the Performance of AlphaEarth Foundation Embeddings for Irrigated Cropland Mapping Across Regions and Years. Remote Sensing, 18, 1065. doi:10.3390/rs18071065.

Zhang, W., Cleary, E., Rowe, F., Chaudhuri, S., Bondarenko, M., Lai, S., Tatem, A. J. (2026). Geospatial foundation-model embeddings improve population estimation unevenly across space and scale. arXiv:2605.01650. doi:10.48550/arXiv.2605.01650.

Zhao, H., Tachet des Combes, R., Zhang, K., Gordon, G. J. (2019). On Learning Invariant Representation for Domain Adaptation. arXiv:1901.09453. doi:10.48550/arXiv.1901.09453.

Zhou, Z., Gao, S., Zhang, Q., Zhang, F., Du, Z. (2026). Quantifying geographic domain shift to decouple the geospatial transferability of human mobility flow generation models. arXiv:2608.21567. doi:10.48550/arXiv.2608.21567.

Zhu, X. X., Xiong, Z., Wang, Y., Stewart, A. J., Heidler, K., Wang, Y., Yuan, Z., Dujardin, T., et al. (2026). On the foundations of Earth foundation models. Communications Earth & Environment, 7. doi:10.1038/s43247-025-03127-x.

Zhuang, Y., Hobeichi, S., Shi, P., Huang, F. (2026). Evaluating AlphaEarth Foundations Embeddings for Wildfire Susceptibility Mapping. arXiv:2608.12663. doi:10.48550/arXiv.2608.12663.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24573054