WHEN LOCAL GAINS FAIL TO TRANSFER: FROZEN EARTH-OBSERVATION EMBEDDINGS ACROSS WILDFIRES

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen Earth-observation embeddings are judged almost entirely by spatially blocked cross-validation inside one study region. We show that this number does not predict accuracy in a new region, we show why, and we show the one setting in which such a model does keep working, using a protocol that needs only a linear probe and labels one already has. The testbed is wildfire, with Copernicus burned-area maps of six fires in Greece and Spain and descriptors from the year before each fire, comparing TESSERA and AlphaEarth with ESA WorldCover classes and annual Sentinel-2 index summaries. Inside a fire, the embeddings identify the burned land 0.05 to 0.13 ROC AUC better than the index summaries, and repeated fold allocations, spatial buffers, a block bootstrap, and gradient-boosted trees leave that margin unchanged. On a fire in another region, they lose 0.15 to 0.18 AUC, and the index summaries lose 0.06, so the three end within a few hundredths of each other. The representation is not the cause. Eight labelled blocks from the new region restore the embedding advantage and give a higher AUC than 59,000 labelled pixels from other regions, and the weight vector fitted in one region is nearly orthogonal to the vector fitted in the others, so the part that carries across regions is small and low-dimensional. Forecasting within a region is a different matter. Fitted on a fire that burned in 2023 and applied to a fire twelve kilometres away that burned in 2024, where nothing used postdates the target fire, TESSERA reaches 0.772 AUC and loses 0.04 against a classifier fitted inside the 2024 fire, while classifiers fitted in other regions lose 0.09 to 0.18. A region with one mapped fire can therefore forecast susceptibility for later fires there; a region without one cannot borrow a model from elsewhere, and every evaluation of a frozen embedding should report a held-out region.

1 INTRODUCTION

Earth-observation foundation models produce per-pixel embeddings that summarise multi-sensor satellite time series into continuous vectors (Lacoste et al., 2023; Jakubik et al., 2023). A large body of work shows that such embeddings serve as reusable features for downstream geospatial tasks (Wang et al., 2022; Rolf et al., 2021; Jean et al., 2019; Tseng et al., 2023; Åström et al., 2025). Such products as TESSERA (Feng et al., 2026; TESSERA Project, University of Cambridge, 2026) and AlphaEarth (Brown et al., 2025; Google and Google DeepMind, 2025) release their embeddings globally at 10 m as fixed annual descriptors. The intended use is to keep the embedding fixed and train a task-specific classifier on top of it. While this use is straightforward, how this mode of use should be evaluated is less settled, with recent work highlighting the lack of standardized evaluation practices for geospatial foundation models (Corley et al.). In this regard, we want to bring two important evaluation criteria into focus: cross-regional and temporal generalization. The first is whether a classifier trained in one region maintains accuracy in another, a step toward a general task-specific model. The second is whether a classifier trained in one region keeps its accuracy there in later years, which is particularly relevant for developing region-specific forecasting models.

In this study, we focus on the task of distinguishing burned from unburned areas following wildfire events by separating the pixels that burned in a wildfire from pixels of the surrounding landscape that did not. To evaluate the cross-regional and temporal dimensions, high-risk wildfire regions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with confirmed fire events were pre-selected and cross-validated against each other. For the temporal dimension, we selected wildfire regions where additional fires occurred in consecutive years. Each fire event is treated as one evaluation unit with both burned and unburned pixels. We took the clean labels from the Copernicus Emergency Management Service (CEMS) Rapid Mapping product (European Commission, Joint Research Centre), a commonly used and publicly available source that supports reproducibility. We use annual embeddings from TESSERA and AlphaEarth as descriptors, selecting the embedding from the year before each fire. We also added two conventional baselines: the ESA WorldCover land-cover class (Zanaga et al., 2022) and annual summaries of three Sentinel-2 vegetation and moisture indices. For the cross-regional evaluation, we study four wildfires from 2023 in three regions, covering two seasons: Rhodes and Evros in Greece, Cáceres in mainland Spain, and Tenerife, a volcanic Atlantic island in Spain. For the temporal evaluation, we also consider two fires in Attica, Greece: Parnitha (2023) and Varnavas (2024), twelve kilometres away. This pair allows us to evaluate transfer one year forward within the same region. We evaluate the pre-fire descriptors using both linear probes and tree-based models.

Our evaluation strategy follows a robust design aimed at assessing generalization across fires, regions, and time. Every descriptor comes from the year before the fire, so no information from the fire itself enters the inputs. Each fire is treated as one evaluation unit, and transfer is evaluated using classifiers trained on entirely different fires. We further examine how much performance can be recovered from a small number of labels from a new fire and compare the weight vectors learned across fires. In the temporal evaluation, both the training labels and descriptors used for the 2024 fire existed before it started, providing a forecasting setting. We find substantially stronger transfer across years within the same landscape than across regions.

With this research, we are contributing in the following way:

- **Cross-regional generalization.** Frozen embeddings consistently outperform conventional descriptors within fires, improving ROC AUC over the Sentinel-2 index baseline by 0.05–0.13, with their combination performing best across all four fires (ROC AUC: 0.832-0.975). This advantage largely disappears across regions, where a small amount of local supervision outperforms substantially larger training sets from other regions, indicating that the embeddings contain useful signal but that their relationship to wildfire occurrence is strongly region-specific.
- **Temporal generalization.** Two fires in Attica, twelve kilometres apart and one year apart, provide a within-region forecasting test. TESSERA retains much of its within-fire performance when transferred from the earlier to the later fire, while AlphaEarth shows a weaker local advantage. Combining the nearby fire with pooled remote fires gives the best transfer for both embeddings, suggesting that within-region performance can persist across years but that spatial and temporal proximity alone is not sufficient.
- **Evaluation protocol.** We introduce a robust protocol for assessing cross-regional and temporal generalization using existing labels and lightweight classifiers. It combines held-out-fire transfer, adaptation with limited local supervision, and comparison of the fitted weight vectors across fires.
- **Data and reproducibility.** We assemble six fires and about 120,000 labelled pixels across four descriptor sets, with fixed spatial fold assignments to ensure reproducible evaluation.

2 Related work

Burned-area mapping from satellite data is a mature field (Chuvieco et al., 2019; Szpakowski & Jensen, 2019; Pettinari & Chuvieco, 2016). Yet what burns within a wildfire is heterogeneous, depending on vegetation, fuel characteristics, and fire conditions (Castilla et al., 2016; Ferster et al., 2016; Kolden & Abatzoglou, 2012; Crowley et al., 2023), and the extent to which burned and unburned areas can be distinguished depends on how the pre-fire landscape is represented. Categorical land-cover products reduce heterogeneous landscapes to discrete classes and can therefore yield different descriptions of the same burned area (Fisher, 1997; Foody, 2002; Comber et al., 2005). In contrast, representation learning for remote sensing has produced task-agnostic embeddings (Jean et al., 2019; Rolf et al., 2021; Tseng et al., 2023; Wang et al., 2022) and benchmarks that evaluate them with frozen heads (Lacoste et al., 2023). TESSERA (Feng et al., 2026) provides global annual embeddings from Sentinel-1 and Sentinel-2 time series, while AlphaEarth Foundations (Brown

et al., 2025) provides global annual embedding fields from optical, radar, and other sources through Earth Engine (Google and Google DeepMind, 2025). Both provide fixed, analysis-ready representations for label-efficient downstream tasks. Recent work has begun to examine how well these representations transfer. Zhuang et al. (2026) use AlphaEarth embeddings for wildfire susceptibility mapping in Victoria and report substantially better transfer to two nearby Australian regions than models based on physical variables. Ko Ko & Atzberger (2026) compare AlphaEarth and TESSERA with Sentinel-1/2 composites for local climate zone mapping across five Swiss cities and examine both cross-city and temporal transfer. Our evaluation differs in its focus on event-level generalization. Every descriptor predates the corresponding fire, each fire is treated as one evaluation unit, and held-out fires are evaluated using probes trained on different fires. This allows us to separate within-fire performance from cross-regional and temporal generalization. Our sampling follows used–available designs in ecology (Johnson, 1980; Northrup et al., 2013; Barbet-Massin et al., 2012), while the spatial validation follows Roberts et al. (2017) and Ploton et al. (2020).

3 METHOD

3.1 DATA

3.1.1 WILDFIRE EVENTS

We use four CEMS Rapid Mapping wildfire activations from 2023 (Table 4, Appendix A). EMSR675 on Rhodes and EMSR686 in Evros are the two large Greek fires of that summer (Copernicus Emergency Management Service, 2023b;d), EMSR667 in Cáceres, Spain, is a spring fire on granite uplands under oak and pine (Copernicus Emergency Management Service, 2023a), and EMSR685 on Tenerife is an August fire in Canary pine forest on volcanic terrain (Copernicus Emergency Management Service, 2023c). For each activation, we take the final delineation product as the mapped burned extent. All 4 fires occurred in 2023, so 2022 is the pre-fire year for every descriptor.

Two further activations serve one experiment only, in Section 4.6. EMSR690 burned the southern slopes of Mount Parnitha in North Attica, Greece, in August 2023 (Copernicus Emergency Management Service, 2023e), and EMSR746 burned the hills around Varnavas, in the same part of Attica, in August 2024 (Copernicus Emergency Management Service, 2024). The two mapped areas lie twelve kilometres apart in the same vegetation, climate and terrain, and the fires are one year apart. The descriptors of the Parnitha fire come from 2022 and those of the Varnavas fire from 2023, so each fire is again described by the year before it burned.

3.1.2 BURNED AND CONTROL PIXEL SELECTION

The unit of analysis is the 10 m TESSERA pixel. Burned pixels are centres inside the burned polygons after a 30 m interior buffer. Control pixels are drawn from the activation area outside the polygons and outside a 500 m exclusion buffer, so they represent the locally available landscape rather than the fire margin. We sample 5,000 burned and 15,000 control pixels per fire with a fixed seed, which gives 80,000 pixels for the four 2023 fires and 40,000 more for the Attica pair. After the exclusions described below, 79,056 pixels enter every analysis of the four fires, with the same comparison mask for every representation. Controls describe the land that was available around the fire, not an estimate of what would have burned under other conditions. Figure 3 (Appendix A) shows the two Greek fires. The Spanish fires are sampled identically.

3.1.3 FOUR DESCRIPTORS OF EACH PIXEL

WorldCover. The ESA WorldCover 2021 v200 class at the pixel, sampled nearest-neighbour and one-hot encoded (Zanaga et al., 2022).

Sentinel-2 index summaries. From all cloud-masked 2022 Level-2A observations (Scene Classification Layer classes 4, 5 and 6 kept) we compute NDVI, NDMI and NBR per observation and summarise each index by its median, interquartile range, 10th and 90th percentiles and valid-observation count, giving 15 features, with the 20 m bands bilinearly resampled to 10 m first. The three count features encode scene coverage rather than the surface, so we report results with and without them.

TESSERA embeddings. The 128-dimensional TESSERA v1.0 annual embedding for 2022 at the pixel (Feng et al., 2026; TESSERA Project, University of Cambridge, 2026). TESSERA stores embeddings as signed 8-bit integers together with one float32 scale per pixel, and the official loader reconstructs the embedding as the product. A per-pixel scale is a per-row multiplier that column standardisation does not remove, so we obtained the scale files for all 58 tiles, use the dequantised embedding throughout, and report the raw integers as a sensitivity variant. 940 control pixels, 902 of them permanent water, have a zero scale and an all-zero integer vector. These all-zero vectors occur only among the controls, so a probe can learn that a missing embedding means a control pixel, which is a fact about the sampling design and not about the land surface. We exclude these pixels from every analysis, which changes any within-fire AUC by at most 0.004.

AlphaEarth embeddings. The 64-dimensional AlphaEarth Foundations annual embedding for 2022 (Brown et al., 2025), read from the Earth Engine Satellite Embedding dataset (Google and Google DeepMind, 2025) at the same pixel centres, nearest pixel at the native 10 m. It comes from a different group, with different inputs and training; its vectors are unit-length floats, so no quantisation question arises, and every sampled pixel has a value. It is a second, independently built frozen embedding, not a controlled variation of the first.

3.2 Experimental design

3.2.1 Probe and within-fire evaluation

All comparisons use L_2-regularised logistic regression with balanced class weights and default regularisation, implemented in scikit-learn (Pedregosa et al., 2011). Continuous features are standardised using training-fold statistics. TESSERA is used in its dequantised form, with the raw integer representation as a sensitivity variant. All descriptors are fixed inputs; only the probe is fitted to the labels. Performance is measured by area under the ROC curve (AUC).

Pixels are grouped into 5 km blocks on the ETRS89-LAEA grid, and five-fold cross-validation holds out complete blocks (Roberts et al., 2017; Ploton et al., 2020). With only 39–90 blocks per fire, the assignment of blocks to folds noticeably affects the result, so we evaluate 21 spatial allocations. Allocation 0 is produced by scikit-learn's `StratifiedGroupKFold` with its default seed. For the remaining 20 allocations, blocks are visited in random order and assigned to the fold with the fewest burned pixels so far. We do not use `shuffle=True`, because it produces little variation in the resulting folds (Section 5). We report the mean and standard deviation of the fold-averaged AUC across allocations, and differences between descriptors are always computed within the same fold before averaging.

Adjacent blocks touch, and the closest test pixel can be only 22–45 m from a training pixel. As a sensitivity analysis, we therefore remove training pixels within 500 m and 1 km of any test pixel. Uncertainty within a fire is estimated by resampling complete blocks with replacement within each test fold using 2,000 bootstrap replicates, without refitting.

3.2.2 Robustness checks

To test whether the results depend on the linear probe, we rerun every descriptor on the same persisted folds using tuned logistic regression and LightGBM gradient-boosted trees (Ke et al., 2017), using the first five of the 21 spatial fold allocations (allocations 0–4). Hyperparameters are selected separately within the training portion of each outer fold using three-fold grouped cross-validation, with identical search spaces for every descriptor set (Appendix B); the outer test fold is never used for model selection. We additionally test the raw versus dequantised TESSERA representation, nodata handling, and the 500 m and 1 km spatial buffers described above.

3.2.3 Generalization to held-out fires

The central generalization test asks what happens when a probe is applied to a fire it has never seen. We evaluate one-to-one transfer by fitting on one fire and testing on each of the other three (12 directed comparisons), and leave-one-fire-out transfer by fitting jointly on three fires and testing on the fourth (four comparisons). Each target fire is scored on the same spatial folds and 21 allocations used for its within-fire evaluation. Transfer loss and differences between descriptors are therefore paired within fold before averaging. For whole-fire transfer AUC, uncertainty is estimated

by resampling spatial blocks with replacement within the target folds using 2,000 bootstrap replicates. Because the regularisation selected within a source fire need not transfer best, we also select the probe regularisation by leaving out one source fire at a time and report this alongside the fixed probe.

Adaptation with target labels. Poor transfer can arise because the representation lacks relevant information for the target fire or because the rule learned elsewhere does not apply there. To separate these cases, we train on the three source fires plus $k \in \{0, 2, 4, 8, 16, 32\}$ labelled blocks from the target fire, and compare this with a probe trained on the same k target blocks alone. Target blocks are sampled from the target's training folds, preserving their spatial separation from the held-out test fold. Recovery with only limited target supervision indicates that the information is present in the representation but is not used correctly by the transferred rule.

Comparing learned rules. A fitted linear probe defines a direction in feature space through its weight vector. For this diagnostic analysis only, features are standardised jointly across the four fires so that the fitted vectors share a common coordinate system. For each target fire, w_t is fitted on four of its folds and s on the other three fires. We decompose $r = w_t - (w_t \cdot s)\, s$, where the component along s captures the direction shared with the other fires and r the target-specific component. The held-out target fold is scored with s, r, and w_t to quantify how much predictive information is carried by the shared and fire-specific directions.

Temporal generalization. The four benchmark fires all occurred in 2023 and therefore cannot separate geographic from temporal generalization. We use two additional fires in Attica, Greece: Parnitha, which burned in August 2023, and Varnavas, twelve kilometres away, which burned in August 2024. Parnitha is represented by 2022 descriptors and Varnavas by 2023 descriptors, so training on Parnitha and applying the probe to Varnavas uses only information available before the target fire and constitutes a forecasting setting.

Varnavas is evaluated on the same spatial folds and with the same paired loss as above. We compare transfer from Parnitha with transfer from each of the four fires in other regions, the four remote fires pooled, and Parnitha combined with the pooled remote fires. We additionally evaluate the reverse transfer from Varnavas to Parnitha and repeat the target-label adaptation experiment on Varnavas with Parnitha as the source.

4 Results

4.1 Frozen embeddings outperform conventional descriptors within fires

Table 1 gives the within-fire AUC for every fire and representation, and Figure 4 in Appendix B shows the spread across allocations for the Greek fires. TESSERA leads the Sentinel-2 summaries by $+0.061$ to $+0.132$ and AlphaEarth by $+0.053$ to $+0.105$, and the paired differences are positive in nearly every allocation. Neither embedding is consistently ahead of the other. TESSERA is ahead on Rhodes by 0.08 in every allocation, the two are level on Evros, and AlphaEarth is ahead on Cáceres and Tenerife by 0.03 and 0.01. Concatenating them gives the highest within-fire AUC on every fire, by $+0.006$ to $+0.047$ over the better single embedding, so locally they carry partly different information. The spread across allocations is not small: on Rhodes, single allocations of TESSERA range from 0.77 to 0.87, and allocation 0, the one the standard splitter produces, sits near the top of that range.

4.2 The within-fire advantage is robust

Spatial buffers and block bootstrap. Removing training pixels within 1 km of every test pixel lowers every representation by a similar amount, so the gap is unchanged, and at that buffer the block bootstrap puts the TESSERA minus Sentinel-2 difference at $+0.111$ (95% interval $+0.014$ to $+0.182$) on Rhodes and $+0.087$ ($+0.060$ to $+0.111$) on Evros (Appendix B).

Classifier family. Under boosting (Table 5, Appendix B), TESSERA still leads the indices by $+0.042$ to $+0.119$ in every fire, in 24 or 25 of 25 folds, and AlphaEarth by $+0.038$ to $+0.103$, so the linear probe is not what holds the index summaries back. An independent re-implementation on

Table 1: Within-fire ROC AUC, logistic regression, mean ± SD over 21 block-to-fold allocations, 79,056 pixels. TESSERA is the dequantised embedding. The lower rows are paired differences computed within fold, with the number of allocations in which the difference is positive. Bold: best representation per fire. Arrows give the direction of improvement.

Representation	Rhodes ↑	Evros ↑	Cáceres ↑	Tenerife ↑
WorldCover	0.629 ± 0.014	0.639 ± 0.011	0.597 ± 0.028	0.814 ± 0.038
Sentinel-2 (15)	0.693 ± 0.028	0.745 ± 0.009	0.763 ± 0.044	0.890 ± 0.030
Sentinel-2, no counts	0.681 ± 0.019	0.742 ± 0.011	0.750 ± 0.047	0.887 ± 0.042
TESSERA	0.825 ± 0.026	0.853 ± 0.007	0.824 ± 0.061	0.958 ± 0.014
AlphaEarth	0.746 ± 0.033	0.850 ± 0.008	0.851 ± 0.035	0.969 ± 0.009
AlphaEarth + TESSERA	$\mathbf{0.832} \pm 0.027$	$\mathbf{0.900} \pm 0.006$	$\mathbf{0.861} \pm 0.032$	$\mathbf{0.975} \pm 0.009$
TESSERA − Sentinel-2 (15)	$+0.132$ (21/21)	$+0.108$ (21/21)	$+0.061$ (20/21)	$+0.068$ (21/21)
AlphaEarth − Sentinel-2 (15)	$+0.053$ (19/21)	$+0.105$ (21/21)	$+0.088$ (21/21)	$+0.079$ (21/21)
AlphaEarth − TESSERA	-0.079 (0/21)	-0.002 (7/21)	$+0.026$ (14/21)	$+0.012$ (19/21)

Table 2: Fire-level summary on matched spatial folds. Within: probe trained inside the fire. Held out: probe trained on the other three fires, scored on the same folds of the same 21 allocations. Loss: within minus held out, paired within fold. S2 = Sentinel-2 index summaries without observation counts. Logistic regression is used here. TESSERA is dequantised. Bold: best of the three representations per fire and column. Arrows give the direction of improvement.

	TESSERA			AlphaEarth			S2 indices		
Held-out fire	within ↑	held out ↑	loss ↓	within ↑	held out ↑	loss ↓	within ↑	held out ↑	loss ↓
Rhodes	**0.825**	0.595	0.230	0.746	0.606	0.140	0.681	**0.632**	**0.049**
Evros	**0.853**	0.674	0.179	0.850	0.681	0.170	0.742	**0.718**	**0.024**
Cáceres	0.824	0.697	0.128	**0.851**	**0.746**	0.105	0.750	0.662	**0.089**
Tenerife	0.958	0.793	0.165	**0.969**	0.793	0.177	0.887	**0.831**	**0.056**
mean	**0.865**	0.690	0.176	0.854	0.707	0.148	0.765	**0.711**	**0.055**

the Greek fires by one of the authors, with XGBoost and pooled AUC, gave the same gap and the same held-out result.

Dequantisation. Using the raw integers instead of the dequantised embedding changes the within-fire AUC by at most 0.006 and the held-out AUC by at most 0.03.

4.3 The embedding advantage disappears across regions

Table 2 and Figure 2 (Appendix B) give, for each fire, the AUC obtained within the fire and the AUC obtained on the same spatial folds when the probe was trained on the other three. TESSERA loses 0.13 to 0.23 AUC, 0.18 on average. AlphaEarth loses 0.11 to 0.18, 0.15 on average. The Sentinel-2 index summaries without observation counts lose 0.02 to 0.09, 0.06 on average. Held out, the three end up in the same place. The mean AUC is 0.69 for TESSERA, 0.71 for AlphaEarth, and 0.71 for the indices. The paired difference between TESSERA and the indices is -0.037, -0.044 and -0.038 on Rhodes, Evros and Tenerife and $+0.035$ on Cáceres, and for AlphaEarth -0.027, -0.037, -0.038 and $+0.084$. The block-bootstrap intervals exclude zero only on Evros and on Cáceres. Concatenating the two embeddings, the best representation within every fire, is held out no better than either alone except on Tenerife. The correct summary is not that the index summaries are ahead in another region, but that the within-fire margin of 0.05 to 0.13 is lost there. The sign of the held-out difference depends on the fire, and on two fires out of four, its size is within the bootstrap uncertainty.

The pairwise designs, gradient boosting, and source-fire tuning give the same result with more noise (Appendix C). Boosting keeps held-out TESSERA far below its boosted within-fire AUC and lowers held-out AlphaEarth on every fire, and choosing the regularisation by leaving out a source fire changes the held-out AUC by at most 0.02.

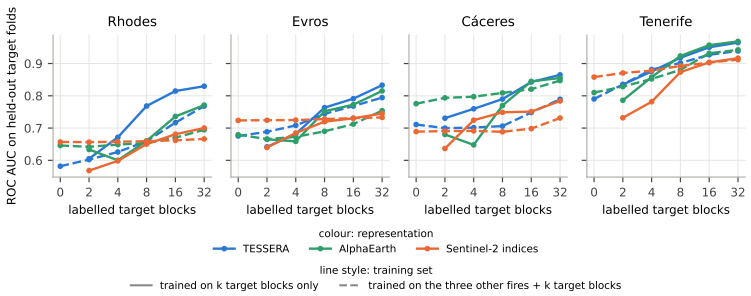


Figure 1: Adaptation curves. AUC on the held-out target fold against the number of labelled target blocks in training, for probes trained on those blocks alone (solid) and on the three other fires plus those blocks (dashed). Mean over five folds and two allocations. The point at zero is the leave-one-fire-out probe.

4.4 Limited local supervision recovers the embedding advantage

Figure 1 shows how much can be recovered from labels from the target fire. With eight labelled blocks of the target, between 11 and 26% of its training blocks, a probe trained on those blocks alone already puts TESSERA ahead of the indices trained the same way on every fire, by $+0.12$ on Rhodes and $+0.04$ on each of the others, and with sixteen blocks the margins are $+0.13$, $+0.06$, $+0.09$ and $+0.05$, most of the full within-fire advantage. AlphaEarth recovers similarly from sixteen blocks. Adding the three other fires, 59,000 labelled pixels, to the same target blocks makes the TESSERA probe worse than the target blocks alone from $k = 8$ on every fire, and from $k = 4$ on Rhodes and Cáceres, by up to 0.10 AUC at $k = 16$. AlphaEarth behaves the same way from $k = 16$, and for the index summaries the other fires help at $k \leq 2$ and are neutral or mildly harmful by $k = 16$. So the embeddings contain what is needed to separate burned from control pixels in every fire. What does not carry from fire to fire is the linear rule, and a rule fitted on other fires actively interferes with one fitted locally.

4.5 The learned linear rule is largely region-specific

Table 6 in Appendix B reports the split of each within-fire direction into the part shared with the other fires and the part specific to that fire. For the embeddings, the weight vector a probe learns inside a fire is nearly orthogonal to the one learned on the other three fires. The cosine is 0.12 to 0.22 for TESSERA and 0.03 to 0.41 for AlphaEarth, and 91 to 99% of the within-fire coefficient norm lies in the fire-specific remainder. That remainder alone reproduces the full within-fire AUC, while the shared direction alone gives the held-out numbers of Table 2. For the index summaries, the shared direction carries most of the within-fire AUC and the specific remainder adds little. The embeddings' local advantage is therefore not a sharper version of a rule shared by all fires. It is a different rule in each fire. The part of the embedding signal that does transfer is also low-dimensional and needs strong regularisation. Held out, a TESSERA probe on the top two to sixteen principal components of the training fires, or one regularised at $C = 10^{-4}$, gives a higher held-out AUC than the full-rank default probe on every fire (Figure 5, Appendix B), and choosing C by leaving out a source fire, which selects 10^{-4} every time, raises the TESSERA held-out mean from 0.69 to 0.73, level with the indices, while within-fire AUC keeps rising to the full 128 dimensions.

4.6 The embedding advantage is retained one year later

We first apply the same within-fire evaluation to the two fires used for the temporal analysis. Across 21 spatial fold allocations, the AUC on Parnitha is 0.908 ± 0.056 for TESSERA, 0.945 ± 0.050 for AlphaEarth, and 0.749 ± 0.090 for the Sentinel-2 index descriptors. On Varnavas, the corresponding values are 0.809 ± 0.092, 0.831 ± 0.078, and 0.795 ± 0.104.

Transfer from Parnitha to Varnavas is strongest for TESSERA (Table 3). It reaches an AUC of 0.772, only 0.040 below the within-Varnavas result, whereas training on a single fire from another region gives 0.628–0.721. AlphaEarth shows a weaker local advantage, with 0.740 from Parnitha

Table 3: Transfer to the Varnavas 2024 fire from different training sources, evaluated on Varnavas's spatial folds (allocation 0, mean of five folds). Loss is the AUC difference relative to a probe trained within Varnavas. Bold indicates the best training source for each descriptor.

	TESSERA		AlphaEarth		S2 indices	
Training source	AUC ↑	loss ↓	AUC ↑	loss ↓	AUC ↑	loss ↓
Rhodes, 2023	0.675	0.137	0.616	0.227	0.757	0.057
Evros, 2023	0.628	0.183	0.697	0.146	**0.790**	**0.025**
Cáceres, 2023	0.629	0.183	0.672	0.171	0.638	0.176
Tenerife, 2023	0.721	0.091	0.758	0.085	0.755	0.059
Parnitha, 12 km away, 2023	0.772	0.040	0.740	0.103	0.649	0.165
The four remote fires pooled	0.765	0.046	0.717	0.126	0.772	0.043
Parnitha and the four pooled	**0.783**	**0.029**	**0.761**	**0.082**	0.771	0.043
within Varnavas	0.812		0.843		0.814	

and 0.758 from Tenerife. The Sentinel-2 index descriptors behave differently, reaching only 0.649 from Parnitha, compared with 0.790 from Evros and 0.757 from Rhodes. Pooling the four remote fires gives 0.765 for TESSERA and 0.717 for AlphaEarth, while combining Parnitha with the pooled fires gives the best transfer result for both embeddings, at 0.783 and 0.761. The adaptation analysis shows the same tendency for TESSERA: adding Parnitha improves performance with few Varnavas labels, whereas adding remote fires becomes harmful from eight target blocks onward (Figure 1).

The reverse transfer is weaker, with losses of 0.29 AUC for TESSERA and 0.16 for the Sentinel-2 index descriptors. Spatial and temporal proximity alone is therefore not sufficient, and the Attica pair should be treated as evidence of temporal transfer within one region rather than as a general property of the embeddings.

5 Discussion

This study evaluated whether frozen Earth observation embeddings can distinguish areas that later burn from those that do not, and how this ability generalizes across fires, regions, and time. Within fires, both embedding products consistently outperform the conventional descriptors: TESSERA exceeds the Sentinel-2 index summaries by 0.061–0.132 AUC and AlphaEarth by 0.053–0.105. Concatenating both embeddings gives the highest within-fire AUC for every fire, with a further gain of 0.006–0.047 over the better single embedding. These results remain stable across spatial allocations, buffering, classifier family, dequantisation, and nodata handling. Across regions, however, this advantage largely disappears. Strong within-fire performance therefore does not imply that a classifier fitted in one region transfers unchanged to another.

Much of the embedding advantage can nevertheless be recovered with limited local supervision. With only eight labelled blocks from the target fire, TESSERA already exceeds the index summaries trained on the same local data by 0.04–0.12 AUC, and with sixteen blocks much of the full within-fire advantage is recovered. Adding approximately 59,000 labelled pixels from the other fires can instead reduce performance relative to using the target blocks alone. This indicates that useful information remains accessible in the embeddings, but that the rule learned in other fires does not transfer well. Consistently, the fitted weight vectors are nearly orthogonal across fires, suggesting that the linear relationship between the embeddings and wildfire occurrence is largely fire-specific. The conventional descriptors show the opposite trade-off: they provide weaker within-fire discrimination, but the relationships learned from them transfer more consistently across regions.

This distinction is important for how frozen Earth observation embeddings are evaluated and used. A benchmark that evaluates a frozen representation only with a locally fitted head measures whether the representation contains information useful in that region, but not whether the fitted rule will generalize elsewhere. Our results suggest that these embeddings may be more useful as reusable representations for lightweight local models than as the basis for a single transferable classifier. This is also an attractive use of foundation models: the expensive representation has already been learned and released, while downstream applications can adapt a small model to local labels instead of training a new representation model for every task and region. This reuse does not by itself

establish a lower environmental cost, but it reduces the need to repeatedly train large representation models from scratch.

The Attica experiment gives an initial indication that such local models may also remain useful over time. For TESSERA, a probe trained on Parnitha in 2023 reaches 0.772 AUC on Varnavas in 2024, only 0.040 below a probe trained within Varnavas itself. Individual fires from other regions reach only 0.628–0.721, while pooling the four remote fires reaches 0.765. AlphaEarth shows a weaker local effect, although combining Parnitha with the pooled remote fires gives the best AlphaEarth transfer result at 0.761. The Attica pair does not establish general temporal stability of the embedding model, but is a first indication that useful cross-year transfer within a region is possible.

This temporal setting also points to a practical direction for future work. In regions with recurrent wildfire risk and previously mapped fires, historical events could provide local supervision while annual foundation-model embeddings provide updated pre-fire descriptors for subsequent years. Region-specific models could then be used to rank areas by susceptibility before a fire occurs and potentially support the prioritisation of monitoring or preventive resources. Because the present study deliberately uses simple probes to test the information contained in the representations, this setting can be extended to richer downstream models without retraining the foundation model itself.

6 Limitations

The study is based on six fires, and the cross-regional fire-level comparisons therefore have a small effective sample size. The temporal analysis is particularly limited because it rests on a single pair of fires and is asymmetric between the two transfer directions. It cannot establish how transfer changes with distance, landscape similarity, or the number of years between source and target fires. A larger evaluation with repeated fires from the same regions over several years is therefore needed before the forecasting result can be generalized.

The fires are restricted to Mediterranean and Macaronesian environments and two consecutive years, and the conventional baseline is intentionally simple. Stronger baselines using additional Sentinel-2 bands, Sentinel-1 radar, terrain, weather, or other physical variables may reduce the observed embedding advantage. The control pixels also differ in composition between fires, so matched or land-cover-stratified controls would help determine how much of the cross-regional loss is compositional. Finally, the fire-specific part of the embedding signal remains unexplained; relating it to land cover, vegetation, terrain, climate, and fuel characteristics is an important next step. Practical sensitivities in grouped cross-validation and nodata handling are documented in Appendix D. Before these models could support operational fire management, they would need to be evaluated across substantially more regions and fire seasons.

7 Conclusion

Using six wildfire events, we evaluated frozen Earth observation embeddings as pre-fire descriptors for distinguishing areas that later burned from those that did not. The embeddings consistently outperform conventional descriptors in within-region evaluation, but this advantage largely disappears under cross-regional generalization. Limited local supervision recovers much of the embedding advantage, indicating that useful information is present in the representations while the fitted linear rule is strongly region-specific. The temporal analysis further suggests that region-specific models can remain useful across years, motivating broader evaluation of temporal generalization for wildfire forecasting.

Reproducibility statement

Every number in the paper is read from a committed output table produced by a script in the accompanying repository. (An anonymised copy of the repository will be provided during the discussion period.) The repository contains the sampling, extraction and analysis code, the 120,000-row pixel table for the six fires with labels, coordinates, block identifiers and all four descriptor sets, the per-pixel TESSERA scale values, the committed fold table for all 21 allocations and four fires, out-of-fold probabilities for every probe, and the environment lock file. CEMS products, WorldCover tiles, and TESSERA tiles are downloaded by a preparation script from their public hosts with recorded

checksums (Appendix A). TESSERA v1.0 tiles are served from the `tessera-embeddings` S3 bucket, since the host we first downloaded them from has been retired. AlphaEarth values are sampled from Earth Engine by a script in the repository, and the sampled values are committed so that the analysis does not require an Earth Engine account.

AI USE STATEMENT

Generative AI tools were used to assist with some language editing as well as improving some code mainly for clarity. All AI-assisted outputs were critically reviewed, tested, and validated by the authors. The authors take full responsibility for the final content of the manuscript, including its code, results, figures, claims, and references.

ETHICS STATEMENT

The study uses only public satellite products and public emergency-mapping polygons at 10 m resolution and involves no personal data. The probes are research models. They have not been validated for operational fire-risk assessment and should not be used for it.

ACKNOWLEDGEMENTS

Will be filled in after the double-blind review process is finished.

REFERENCES

Oskar Åström, Simon Månsson, Isac Lazar, Magnus Nilsson, Joakim Ekelöf, Andreas Oxenstierna, and Alexandros Sopasakis. Predicting intra-field yield variations for winter wheat using remote sensing and graph attention networks. *Computers and Electronics in Agriculture*, 237: 110499, 2025. doi: 10.1016/j.compag.2025.110499. URL `https://doi.org/10.1016/j.compag.2025.110499`.

Morgane Barbet-Massin, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2):327–338, 2012. doi: 10.1111/j.2041-210X.2011.00172.x.

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. *arXiv preprint arXiv:2507.22291*, 2025. doi: 10.48550/arXiv.2507.22291. URL `https://arxiv.org/abs/2507.22291`.

Guillermo Castilla, Sebastien Rodrigue, Rob S. Skakun, and Ron J. Hall. Four National Maps of Broad Forest Type Provide Inconsistent Answers to the Question of What Burns in Canada. *Remote Sensing*, 8(7):539, 2016. doi: 10.3390/rs8070539.

Emilio Chuvieco, Florent Mouillot, Guido R. van der Werf, Jesús San Miguel, Mihai Tanase, Nikos Koutsias, Mariano García, Marta Yebra, Marc Padilla, Ioannis Gitas, Angelika Heil, Todd J. Hawbaker, and Louis Giglio. Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sensing of Environment*, 225:45–64, 2019. doi: 10.1016/j.rse.2019.02.013.

Alexis Comber, Peter Fisher, and Richard Wadsworth. What is Land Cover? *Environment and Planning B: Planning and Design*, 32(2):199–209, 2005. doi: 10.1068/b31135.

Copernicus Emergency Management Service. EMSR667: Wildfire in Cáceres, Spain. Copernicus Emergency Management Service, Rapid Mapping activation EMSR667, 2023a.

Copernicus Emergency Management Service. EMSR675: Wildfire on Rhodes island, Greece. Copernicus Emergency Management Service, Rapid Mapping activation EMSR675, 2023b.

Copernicus Emergency Management Service. EMSR685: Fire in Tenerife, Spain. Copernicus Emergency Management Service, Rapid Mapping activation EMSR685, 2023c.

Copernicus Emergency Management Service. EMSR686: Wildfire in East Macedonia, Greece. Copernicus Emergency Management Service, Rapid Mapping activation EMSR686, 2023d.

Copernicus Emergency Management Service. EMSR690: Wildfire in north attica, greece. Copernicus Emergency Management Service, Rapid Mapping activation EMSR690, 2023e.

Copernicus Emergency Management Service. EMSR746: Wildfire in attica region, greece. Copernicus Emergency Management Service, Rapid Mapping activation EMSR746, 2024.

Isaac Corley, Nils Lehmann, Caleb Robinson, Gabriel Tseng, Anthony Fuller, Hamed Alemohammad, Evan Shelhamer, Jennifer Marcus, and Hannah Kerner. No One Knows the State of the Art in Geospatial Foundation Models. URL https://arxiv.org/abs/2605.12678.

Morgan A. Crowley, Christopher A. Stockdale, Joshua M. Johnston, Michael A. Wulder, Tianjia Liu, Jessica L. McCarty, Jesse T. Rieb, Jeffrey A. Cardille, and Joanne C. White. Towards a whole-system framework for wildfire monitoring using Earth observations. *Global Change Biology*, 29 (6):1423–1436, 2023. doi: 10.1111/gcb.16567.

European Commission, Joint Research Centre. Copernicus Emergency Management Service Rapid Mapping. JRC Data Catalogue, collection CEMS-RM.

Z. Feng, C. Atzberger, S. Jaffer, J. Knezevic, S. Sormunen, R. Young, M. C. Lisaius, M. Immitzer, T. Jackson, J. Ball, D. A. Coomes, A. Madhavapeddy, A. Blake, and S. Keshav. TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34818–34831, 2026.

Colin J. Ferster, Bianca N. I. Eskelson, David W. Andison, and Valerie M. LeMay. Vegetation Mortality within Natural Wildfire Events in the Western Canadian Boreal Forest: What Burns and Why? *Forests*, 7(9):187, 2016. doi: 10.3390/f7090187.

P. Fisher. The pixel: A snare and a delusion. *International Journal of Remote Sensing*, 18(3): 679–685, 1997. doi: 10.1080/014311697219015.

Giles M. Foody. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201, 2002. doi: 10.1016/S0034-4257(01)00295-4.

Google and Google DeepMind. Google satellite embedding V1, annual (GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL). Earth Engine Data Catalog, 2025. URL https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_SATELLITE_EMBEDDING_V1_ANNUAL. 64-band annual embedding fields at 10 m, 2017–2024, CC BY 4.0.

J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, K. Ankur, M. Ramasubramanian, I. Gurung, S. Loughlin, M. Ahmadi, F. Kordi, H. Alemohammad, M. Maskey, R. Ganti, K. Weldemariam, and R. Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence. arXiv:2310.18660, 2023.

Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3967–3974. AAAI, 2019. doi: 10.1609/aaai.v33i01.33013967.

Douglas H. Johnson. The Comparison of Usage and Availability Measurements for Evaluating Resource Preference. *Ecology*, 61(1):65–71, 1980. doi: 10.2307/1937156.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Htet Yamin Ko Ko and Clement Atzberger. Exploring the potential of AlphaEarth and TESSERA embeddings for fine-scale local climate zone mapping: A case study across five cities in Switzerland. *arXiv preprint arXiv:2606.20034*, 2026. URL https://arxiv.org/abs/2606.20034.

Crystal A. Kolden and John T. Abatzoglou. Wildfire Consumption and Interannual Impacts by Land Cover in Alaskan Boreal Forest. *Fire Ecology*, 8(1):98–114, 2012. doi: 10.4996/fireecology.0801098.

A. Lacoste, N. Lehmann, P. Rodriguez, E. D. Sherwin, H. Kerner, B. Lütjens, J. Irvin, D. Dao, H. Alemohammad, A. Drouin, M. Gunturkun, G. Huang, D. Vazquez, D. Newman, Y. Bengio, S. Ermon, and X. X. Zhu. GEO-Bench: Toward Foundation Models for Earth Monitoring. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Joseph M. Northrup, Mevin B. Hooten, Charles R. Anderson Jr., and George Wittemyer. Practical guidance on characterizing availability in resource selection functions under a use–availability design. *Ecology*, 94(7):1456–1463, 2013. doi: 10.1890/12-1688.1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

M. Lucrecia Pettinari and Emilio Chuvieco. Generation of a global fuel data set using the Fuel Characteristic Classification System. *Biogeosciences*, 13(7):2061–2076, 2016. doi: 10.5194/bg-13-2061-2016.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11:4540, 2020. doi: 10.1038/s41467-020-18321-y.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12:4392, 2021. doi: 10.1038/s41467-021-24638-z.

David Szpakowski and Jennifer Jensen. A Review of the Applications of Remote Sensing in Fire Ecology. *Remote Sensing*, 11(22):2638, 2019. doi: 10.3390/rs11222638.

TESSERA Project, University of Cambridge. Temporal Embeddings of Surface Spectra for Earth Representation and Analysis (TESSERA). Software repository, 2026. URL https://github.com/ucam-eo/tessera.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv:2304.14065, 2023.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-Supervised Learning in Remote Sensing: A Review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022. doi: 10.1109/MGRS.2022.3198244.

D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. ESA WorldCover 10 m 2021 v200. Zenodo, 2022. URL https://doi.org/10.5281/zenodo.7254221.

Yuan Zhuang, Sanaa Hobeichi, Peng Shi, and Fei Huang. Evaluating AlphaEarth Foundations embeddings for wildfire susceptibility mapping. *arXiv preprint arXiv:2608.12663*, 2026. URL https://arxiv.org/abs/2608.12663.

A Data and processing details

Table 4: The six wildfire events. The first four are the benchmark. The last two are the Attica pair of Section 4.6, with descriptors from 2022 and 2023 respectively. Blocks are 5 km spatial cross-validation blocks. No-data pixels are control pixels with no TESSERA embedding (all-zero vector), nearly all permanent water. They are excluded from all analyses. Sentinel-2 observations are the median number of cloud-free pre-fire observations per pixel.

Activation	Fire	Date	CEMS product	Burned (ha)	Blocks	Nodata	S2 obs.
EMSR675	Rhodes, Greece	Jul 2023	DEL_MONIT08	17,770	48	440	122
EMSR686	Evros, Greece	Aug 2023	DEL_MONIT09	93,939	90	338	122
EMSR667	Cáceres, Spain	May 2023	DEL_PRODUCT v2	9,595	47	0	228
EMSR685	Tenerife, Spain	Aug 2023	DEL_MONIT02	10,960	39	162	58
EMSR690	Parnitha, Greece	Aug 2023	DEL_MONIT03	6,190	29	0	148
EMSR746	Varnavas, Greece	Aug 2024	DEL_MONIT02	10,405	20	0	128

CEMS products. EMSR675 `EMSR675_AOI01_DEL_MONIT08_v1`, EMSR686 `EMSR686_AOI01_DEL_MONIT09_v1`, EMSR667 `EMSR667_AOI01_DEL_PRODUCT_v2`, EMSR685 `EMSR685_AOI01_DEL_MONIT02_v1`, EMSR690 `EMSR690_AOI01_DEL_MONIT03_v1` and EMSR746 `EMSR746_AOI01_DEL_MONIT02_v1`. The `observedEventA` layer gives burned polygons, and `areaOfInterestA` the activation area from which controls are drawn. Two of the 112 Evros polygons were topologically invalid on read and were repaired without change of area.

Sampling. Interior buffer 30 m, exclusion buffer 500 m, control ratio 3, caps 5,000 and 15,000, seed 42. Blocks are 5 km cells of the EPSG:3035 grid. Pixels are deduplicated by tile ownership at TESSERA tile boundaries.

Sentinel-2. Level-2A from the Microsoft Planetary Computer STAC, all scenes of the pre-fire year over each activation bounding box (438 Rhodes, 334 Evros, 635 Cáceres, 99 Tenerife, 337 Parnitha, 234 Varnavas). Reflectance is $(\mathrm{DN} - 1000)/10000$. B04 and B08 nearest at 10 m, B11 and B12 bilinear from 20 m. Pixels with no valid observation (four on Tenerife) are dropped.

TESSERA. v1.0, variant `vultr`, year 2022 for the five 2023 fires and year 2023 for the Varnavas fire, 58 tiles for the benchmark and 19 more for the Attica pair. Embedding $= \mathrm{int8} \times \mathrm{scale}$, where scale is a float32 per pixel in the range 0.04 to 0.12 over our pixels. Fold seeds and classifier settings are in the repository.

AlphaEarth. Earth Engine image collection `GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL`, bands A00 to A63, sampled at each pixel centre at 10 m without resampling, for the pre-fire year of each activation. Values lie in $[-0.48, 0.50]$ and every sampled vector has unit norm. The dataset is released under CC BY 4.0 and is produced by Google and Google DeepMind.

B Additional within-fire results

Classifier grids. The tuned logistic regression searches $C \in \{0.01, 0.1, 1, 10\}$. LightGBM uses 300 trees, learning rate 0.05, leaves $\in \{15, 63\}$, minimum leaf size $\in \{20, 100\}$ and balanced weights. Boosting lifts the index summaries by $+0.005$ to $+0.036$, TESSERA by $+0.003$ to $+0.033$ and AlphaEarth by $+0.008$ to $+0.031$, and on the held-out set it reaches 0.63, 0.65, 0.69 and 0.86 for TESSERA against 0.60, 0.70, 0.65 and 0.85 for the indices.



Figure 2: AUC on each fire's own spatial folds when the probe is trained within the fire (filled) and when it is trained on the other three fires (open), for TESSERA, AlphaEarth, and the Sentinel-2 index summaries.

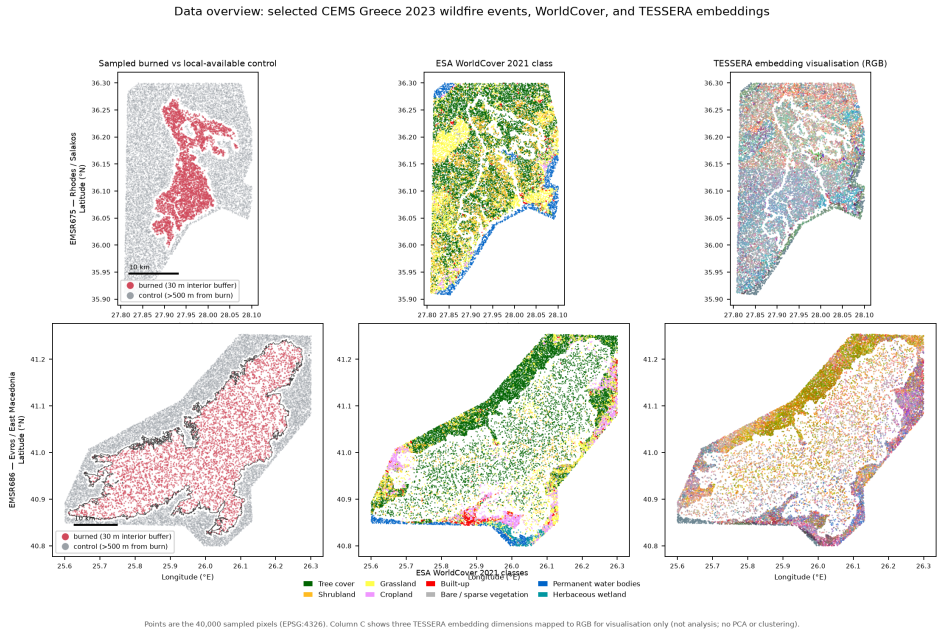


Figure 3: Sampled burned and control pixels, WorldCover class, and a three-dimensional TESSERA visualisation for Rhodes (top) and Evros (bottom).

Buffers. Removing training pixels within 1 km of every test pixel, about 14% of the training set, lowers TESSERA on allocation 0 of the Greek fires from 0.846 to 0.820 (Rhodes) and from 0.855 to 0.839 (Evros), with Sentinel-2 at 0.709 and 0.753 under the same buffer. Held out, dequantisation helps by $+0.02$ on Evros and $+0.03$ on Tenerife, with bootstrap intervals excluding zero, and by less than 0.01 elsewhere.

Tenerife is the easiest fire, and that is informative. Tenerife has the highest AUC for every representation, including 0.81 for the land-cover class alone. Its burned area is almost entirely pine forest, and its controls are a mixture of shrub, bare lava, built-up land, and coast, so the burned–control contrast is largely a land-cover contrast. It is also the easiest target for probes trained on Evros (0.75 to 0.85 for every representation) and the worst source (0.49 to 0.71 on the other fires). The transfer numbers reflect how alike the burned–control contrasts of two fires are, not how much a representation knows about fire.

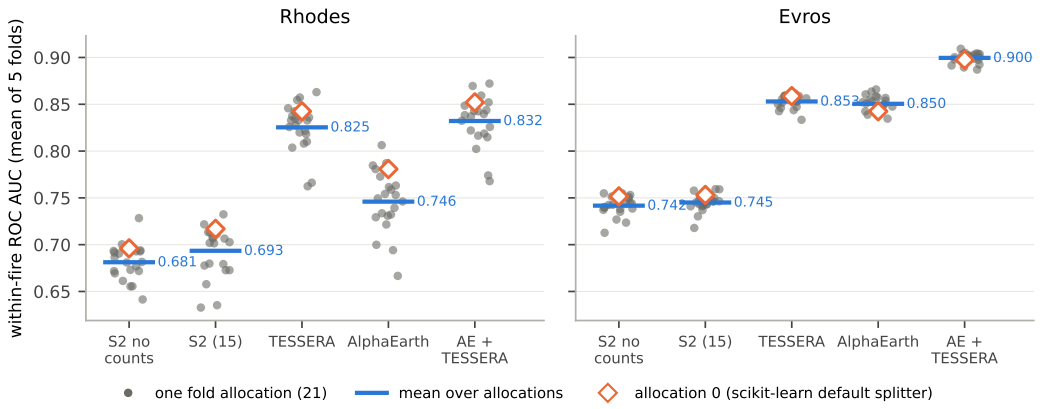


Figure 4: Within-fire AUC under each of 21 block-to-fold allocations for the two Greek fires. Grey points are single allocations, bars are their mean, and diamonds are allocation 0, the one produced by scikit-learn's default splitter.

Table 5: Within-fire ROC AUC with a tuned logistic regression (LR) and with gradient boosting (GBM) on the same persisted folds (allocations 0 to 4, mean over folds and allocations, TESSERA raw integers, which differ from the dequantised embedding by at most 0.006 within fire). The last two columns are the paired TESSERA minus Sentinel-2 (15) difference under each classifier. Bold: best representation per fire under each classifier.

	Sentinel-2 (15)		S2 no counts		TESSERA		AlphaEarth		Δ (T − S2)	
Fire	LR ↑	GBM ↑	LR ↑	GBM ↑	LR ↑	GBM ↑	LR ↑	GBM ↑	LR	GBM
Rhodes	0.700	0.705	0.691	0.706	**0.821**	**0.824**	0.734	0.743	+0.120	+0.119
Evros	0.748	0.784	0.748	0.759	**0.848**	0.858	0.846	**0.877**	+0.100	+0.074
Cáceres	0.774	0.791	0.758	0.767	0.823	0.856	**0.866**	**0.894**	+0.048	+0.064
Tenerife	0.891	0.926	0.885	0.923	0.955	0.968	**0.968**	**0.976**	+0.065	+0.042

C Full transfer results

Pairwise transfer, boosting and source-fire tuning. The pairwise designs give the same result as the leave-one-fire-out designs, with more noise. Across the twelve directions, the indices are ahead of TESSERA in five and behind in seven, and ahead of AlphaEarth in eight. The largest gaps in either direction involve Tenerife as source or target (Cáceres to Tenerife gives TESSERA 0.68 and the indices 0.53. Rhodes to Tenerife gives the indices 0.83 and AlphaEarth 0.59). Going from one training fire to three raises the mean held-out AUC of TESSERA by about 0.01 and of AlphaEarth by about 0.08 and does not restore the within-fire margin, though four fires cannot say what many would do. Choosing the probe's regularisation by leaving out a source fire changes the held-out AUC by at most 0.02 in either direction. Gradient boosting changes the order of the representations but not the conclusion. Held out, boosted TESSERA reaches 0.63, 0.65, 0.69 and 0.86 against 0.60, 0.70, 0.65 and 0.85 for the boosted indices, ahead on three fires by 0.01 to 0.04 but far below its boosted within-fire AUC of 0.82 to 0.97, while boosted AlphaEarth falls to 0.55, 0.66, 0.68 and 0.81, below its linear probe on every fire. One failure is worth explaining. The full Sentinel-2 set, including observation counts, trained with boosting on the other fires, scores 0.34 on Cáceres, well below chance, because Cáceres has twice the cloud-free coverage of the Greek fires and the trees split on the count columns. Observation counts are a property of the acquisition, not the surface, and should not be in a baseline that is meant to transfer.

The Attica pair. Whole-fire AUC, which ranks all pixels of the target together instead of averaging over folds, gives the same picture as Table 3. From Parnitha to Varnavas, TESSERA reaches 0.738, AlphaEarth 0.719, the index summaries 0.669, and the two embeddings concatenated 0.754. In the reverse direction, from Varnavas to Parnitha, TESSERA reaches 0.619, AlphaEarth 0.695, and

Table 6: Within-fire probe direction decomposed along the direction learned on the other three fires (allocation 0, mean of five folds, global standardisation). cos: cosine between the two directions. AUC on the test fold using the shared direction alone, the fire-specific remainder alone, and the full within-fire direction. Bold: larger of shared and specific.

	TESSERA				AlphaEarth				S2 indices			
Fire	cos	shared	specific	full	cos	shared	specific	full	cos	shared	specific	full
Rhodes	0.12	0.594	**0.837**	0.843	0.20	0.605	**0.770**	0.781	0.54	**0.634**	0.625	0.697
Evros	0.20	0.674	**0.847**	0.859	0.41	0.678	**0.747**	0.843	0.24	0.719	**0.746**	0.751
Cáceres	0.22	0.699	**0.859**	0.869	0.03	0.752	**0.868**	0.870	−0.25	0.676	**0.765**	0.767
Tenerife	0.16	0.817	**0.955**	0.957	0.09	0.793	**0.957**	0.959	0.71	**0.855**	0.831	0.912

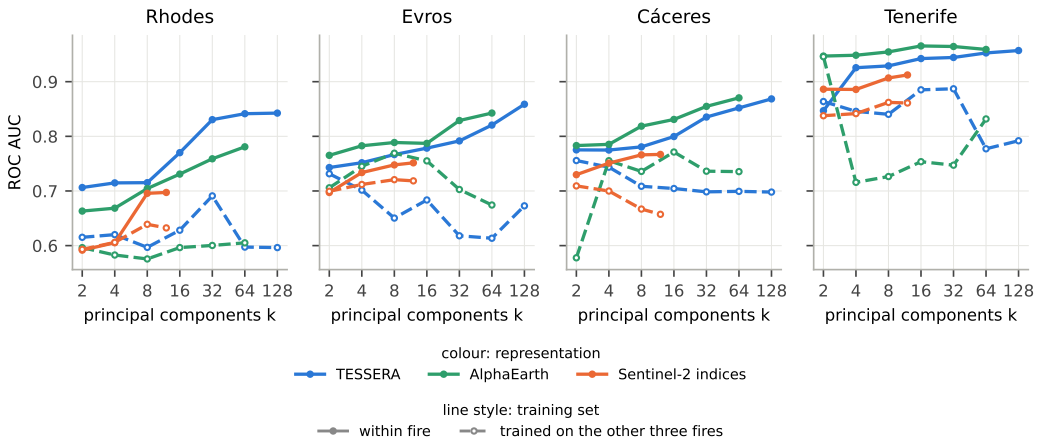


Figure 5: AUC of the probe on the top k principal components, within-fire (PCA fitted on the training folds) and held-out (PCA fitted on the three training fires), allocation 0. Within-fire AUC of the embeddings keeps rising to full rank. Held-out AUC peaks at a few components.

the index summaries 0.593, against within-fire values of 0.91, 0.95, and 0.75 on Parnitha. The embeddings therefore stay ahead of the index summaries in both directions of this pair, which happens for no remote pair, but only the forward direction keeps most of the within-fire performance.

D Two pitfalls we met - beyond this study

We also report two practical problems we encountered and which we believe could be valuable to know for the community beyond this study alone. First, scikit-learn's `StratifiedGroupKFold` only shuffles groups that have similar class ratios, as its documentation states. With a few dozen blocks per fire, almost every random seed then produces the same split into folds, and for Rhodes, 21 seeds gave only 3 distinct splits, ten of them identical, so a standard deviation computed over seeds is not a measure of split-to-split variability. We drew our own random block-to-fold allocations instead, saved all 21 of them in the repository, and used that saved table for every experiment. Second, TESSERA marks pixels without an embedding by an all-zero vector. In our sample, these are almost all permanent water, so they occur only among the controls, and a classifier can learn "all zeros means control" without learning anything about the land surface. The effect here was at most 0.004 AUC, but it grows with class imbalance, and a cross-validation that leaves such pixels in cannot detect it.

E Descriptive results for the Greek fires

Burned pixels in the Greek fires are over-represented in tree cover (0.683 of burned versus 0.517 of control pixels) and shrubland (0.162 versus 0.087) and under-represented in grassland and cropland.

Table 7: Pairwise and leave-one-fire-out transfer ROC AUC over the whole target fire, logistic regression, nodata excluded. Whole-fire AUCs differ from the fold-averaged values of Table 2 because they also rank pixels across folds. Bold: best representation per row. S2 (15) includes observation counts and S2 idx excludes them. T. deq. = TESSERA dequantised, AE = AlphaEarth, AE+T = AlphaEarth and TESSERA concatenated.

Train	Test	S2 (15) ↑	S2 idx ↑	TESSERA ↑	T. deq. ↑	AE ↑	AE+T ↑
Rhodes	Evros	0.640	**0.663**	0.654	0.642	0.567	0.639
Rhodes	Cáceres	0.526	0.604	0.620	0.617	0.684	**0.732**
Rhodes	Tenerife	**0.847**	0.828	0.777	0.786	0.589	0.678
Evros	Rhodes	0.562	**0.632**	0.568	0.563	0.553	0.495
Evros	Cáceres	**0.717**	0.702	0.676	0.654	0.663	0.684
Evros	Tenerife	0.847	0.849	0.753	0.766	0.826	**0.873**
Cáceres	Rhodes	0.489	0.572	0.582	0.596	0.606	**0.641**
Cáceres	Evros	0.682	0.678	0.681	**0.694**	0.621	0.689
Cáceres	Tenerife	0.511	0.527	0.684	**0.732**	0.582	0.648
Tenerife	Rhodes	0.524	0.626	0.650	**0.653**	0.494	0.546
Tenerife	Evros	0.689	0.706	0.686	**0.712**	0.666	0.688
Tenerife	Cáceres	0.683	0.678	0.689	0.696	0.678	**0.700**
Evros + Cáceres + Tenerife	Rhodes	**0.660**	0.632	0.595	0.596	0.605	0.606
Rhodes + Cáceres + Tenerife	Evros	**0.719**	**0.719**	0.655	0.673	0.675	0.671
Rhodes + Evros + Tenerife	Cáceres	0.629	0.658	0.700	0.699	0.735	**0.741**
Rhodes + Evros + Cáceres	Tenerife	0.864	0.861	0.766	0.792	0.832	**0.915**

Table 8: Leave-one-fire-out transfer ROC AUC with gradient boosting, tuned inside the training fires. Bold: best representation per row.

Held-out fire	S2 (15) ↑	S2 idx ↑	TESSERA ↑	TESSERA deq. ↑	AlphaEarth ↑
Rhodes	**0.645**	0.601	0.630	0.632	0.550
Evros	0.638	**0.697**	0.651	0.668	0.658
Cáceres	0.342	0.648	**0.687**	0.681	0.680
Tenerife	0.798	0.852	0.859	**0.861**	0.806

In the 128-dimensional embedding space, the burned and control centroids are separated (Euclidean distance 76.9 pooled, on raw integer values), and the separation persists within every WorldCover class with at least 100 burned and 100 control pixels, so the embedding contrast is not a restatement of land-cover composition. The full tables are in the repository.