VEGBENCH: EVALUATING THE ECOLOGICAL RESOLUTION OF GEOSPATIAL MODELS WITH EXPERT FIELD SURVEYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Categorizing ecosystems is an important component of both ecological modeling and conservation decision-making. We introduce VEGBENCH, the first geospatial benchmark to focus on how well models recover hierarchical categorization of ecosystems, from coarse groupings (i.e. forest vs. desert) down to fine-scale vegetation types useful for diverse ecological tasks like protecting threatened species and mitigating wildfire. VEGBENCH standardizes decades of expert-collected vegetation classification and accuracy assessment field surveys across the state of California, spanning 42,393 surveys of 387 vegetation alliances. We utilize VEGBENCH to compare modern geoFMs against remote sensing and environmental covariate baselines, and find that even the best current models can capture broad ecological structure but fall short as ecological resolution approaches the fine-grained distinctions required for conservation decision-making. VEGBENCH offers the geospatial AI community a much needed testbed to assess whether and when models capture *ecologically-meaningful* variation.

1 INTRODUCTION

From guiding prescribed burns (Finney, 1998) to delineating endangered species habitat (CDFW, 2010), categorizing ecosystems is an important component of ecological modeling (Thuiller et al., 2005), ecosystem management (Rollins, 2009), and conservation decision-making (CDFW, 2025). Many of these tasks require habitats to be characterized to a fine level of spatial and ecological granularity. Within even a single kilometer, ecosystems may differ substantially in species composition and vegetation structure (Fig. 1, Woodcock & Strahler (1987); CDFW & CNPS (2014)). To help standardize complex, fine-scale ecosystem categorization across factors like soil, climate and topography (Faber-Langendoen et al., 2014), the ecology community has converged on *vegetation classification* as a systematic framework (Mucina, 1997) (see A.1 for term definitions).

Vegetation classification involves categorizing local habitats into *vegetation associations*: recurring groupings of plant species that co-occur under similar environmental conditions (Clements, 1920). Similar to species taxonomy, these vegetation associations are then organized into a *vegetation hierarchy*, spanning from broad macroecological categories like formation classes down to fine-scale communities like *alliances* defined by characteristic species composition and vegetation structure (Fig. 1a, Franklin et al. (2012); Jennings et al. (2009)). Vegetation alliances and associations are much more fine-grained than most geospatial classification tasks, e.g. land use land cover (LULC) classification. To whit, the U.S. National Vegetation Classification (USNVC) contains over 7,000 vegetation associations and more than 1,500 alliances across the U.S. compared to 13 broad LULC categories (Fig. 1b, FDGC (2008)). Expert ecologists collect extensive ground-validated field surveys in order to classify and verify these vegetation alliances, as they are often not possible to confirm to the finest scale visually from remotely sensed imagery (Fig. 1c, Sawyer et al. (2009)).

Models that learn representations of the Earth from large collections of satellite imagery and other geospatial data—often termed Geospatial Foundation Models (geoFMs)—pose a promising avenue to improve the vegetation classification process. GeoFMs are increasingly used as general-purpose covariates for environmental prediction (Klemmer et al., 2025b; Brown et al., 2025), and have been applied to species occupancy modeling, habitat classification, and other ecological prediction tasks

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.

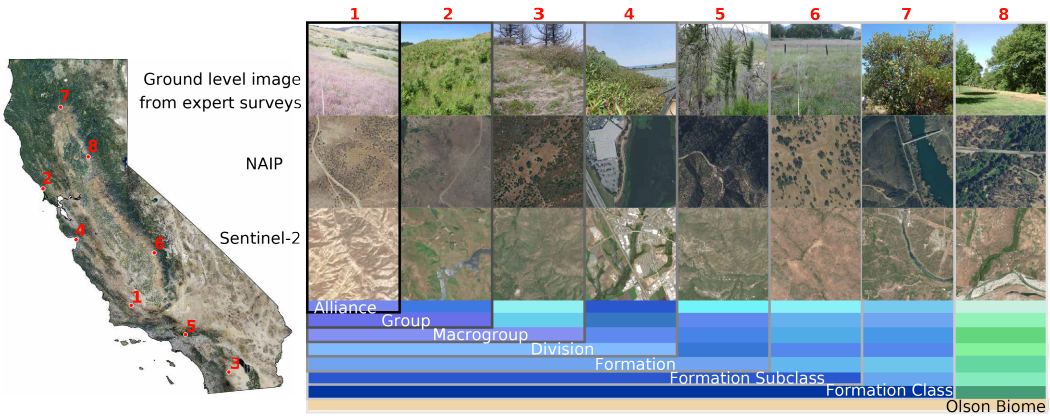

Figure 1: **Vegetation classification is a unified way of categorizing ecosystems worldwide.** Vegetation classification categorizes ecosystems by their unique vegetation (Clements, 1920), and is a common tool for delineating areas of high conservation management priority (CDFW, e). Vegetation classification is **(1)** hierarchical, mirroring the tree of life (Faber-Langendoen et al., 2025); **(2)** fine-scale, reaching tens of thousands of classes (USNVC, 2025; EEA, 2021); and **(3)** expert-driven, with vegetation types identified and validated on the ground by regional experts (Sawyer et al., 2009). Most geospatial benchmarks test ecosystem categorization ability using coarse, non-hierarchical classifications like Olson biomes (Olson et al., 2001).

Table 1: **VEGBENCH fills a geospatial benchmarking gap** and provides a new ecologically-centered, high-granularity task suite for measuring ecosystem categorization capability.

Benchmark / Evaluation Data	Ecosystem classification	Hierarchical	Ground-validated field data	Accuracy assessment
US Trees (Brown et al., 2025)	×	×	✓	×
PANGAEA (Marsocci et al., 2025)	×	×	×	×
GEO-Bench (Lacoste et al., 2023)	×	×	2/12 tasks	×
GEO-Bench-2 (Simumba et al., 2026)	×	1/19 tasks	1/19 tasks	×
Olson Biome (Olson et al., 2001)	✓	×	×	×
VEGBENCH (ours)	✓	✓	✓	✓

(Haucke et al., 2025; Si-Moussi et al., 2025; Kay et al., 2026; Ishikawa et al., 2026). However, standard geospatial benchmarks mostly focus on broad LULC classification (Lacoste et al., 2023; Marsocci et al., 2025; Simumba et al., 2026) or coarse biome prediction (Klemmer et al., 2025a; Sastry et al., 2025; 2026; Rao et al., 2026; Joseph et al., 2026), both of which are insufficient for fully characterizing model capability for fine-grained ecosystem characterization (Table 1). Furthermore, many geospatial benchmark task evaluation labels are themselves extracted from remote sensing-based models or photointerpreted maps rather than field-collected observations (Stanimirova et al., 2023; USDA-NASS, 2025; Brown et al., 2020). We need a benchmark that reliably tests whether geoFMs can systematically categorize ecosystems down to fine-scale ecological resolutions.

Here we utilize tens of thousands of standardized, expert-collected field surveys from California's Vegetation Classification and Mapping Program (VegCAMP) to develop **VEGBENCH**, the first benchmark that explores the capability horizon of geoFMs and other models across the full vegetation hierarchy (Sawyer et al., 2009; CDFW, c; CDFW & CNPS, 2025; CDFW, a). The state of California is a valuable testbed for this challenging task of vegetation classification: it is the most ecologically-diverse in the United States, and has invested more than USD$10M in the collection and standardization of vegetation data over the last 50 years (Ratchford et al., 2026), providing the foundation of our curated evaluation.

Our key contributions are:

- **VEGBENCH: the first large-scale, hierarchical vegetation classification benchmark.** Built from 42,393 expert-collected VegCAMP field surveys across 387 vegetation alliances, VEGBENCH spans multiple levels of the USNVC hierarchy and includes independently-collected and expert-verified accuracy assessment surveys for ground-validated evaluation.
- **We evaluate a diverse set of models for vegetation classification at varying levels of granularity**. Spanning geoFMs to simple baselines, we analyze ML model performance across the MCV vegetation hierarchy and classifications similar to existing benchmarks.
- **We reveal a significant expert-to-geoFM performance gap.** Comparing vegetation classification accuracy to expert maps generated by VegCAMP's vegetation ecologists, we show that ML model performance falls overwhelmingly short of expert map accuracy.
- **We characterize the limits of current geoFMs for fine-scale vegetation classification.** We find that model performance deteriorates at the finest-grained vegetation categories, revealing a substantial gap between current geoFM representations and the requirements of operational vegetation classification and conservation decision-makers' needs.

2 RELATED WORK

Existing geospatial benchmarks. Recent geospatial benchmarks (Table 1) have sought to standardize geospatial foundation model evaluation across a broad range of Earth observation tasks (Lacoste et al., 2023; Marsocci et al., 2025; Simumba et al., 2026), including land use/land cover, crop classification, segmentation, and, in some cases, ecological targets such as tree-species or biome classification. For example, GEO-Bench includes TreeSatAI alongside conventional land-cover tasks (Lacoste et al., 2023), while several geospatial evaluations use Olson biomes as a broad ecosystem target (Klemmer et al., 2025a; Sastry et al., 2025; 2026; Rao et al., 2026). These data vary substantially in how they are produced: some are map products generated from remote sensing-based ML models or photointerpretation (Stanimirova et al., 2023; USDA-NASS, 2025; Brown et al., 2020), while other like LUCAS (d'Andrimont et al., 2020), CropHarvest (Tseng et al., 2021), and U.S. Trees (Brown et al., 2025) use ground-verified data collections for validation.

Ecological classification and mapping. Broad ecoregional frameworks, such as U.S. EPA ecoregions (Omernik & Griffith, 2014) partition landscapes using recurring combinations of climate, landform, soils, hydrology, and biota (Smith et al., 2018), while the IUCN Global Ecosystem Typology provides a global hierarchy spanning broad functional groups to finer ecosystem types (Keith et al., 2020). Vegetation-focused frameworks such as EcoVeg and the USNVC instead emphasize plant-community composition and structure (Faber-Langendoen et al., 2018; Baldwin et al., 2019). Regional systems including South Africa's National Vegetation Map and Europe's Nature Information System (EUNIS) similarly provide standardized ecological units for biodiversity assessment and conservation planning (Dayaram et al., 2019; Chytrý et al., 2020). Operational mapping programs such as LANDFIRE, CALVEG, and CWHR (see A.2 for acronym definitions) translate related ecological concepts into spatial products used for vegetation, wildlife, and wildfire management (Rollins, 2009; USDA, Forest Service; CDFW, b). While some attempts at geoFM-based habitat and biome mapping have recently emerged (Si-Moussi et al., 2025; Joseph et al., 2026), the field data that underpin these efforts are broadly underutilized in geospatial benchmarking.

Ecological prediction with geospatial representations. Ecological modelers are increasingly utilizing geoFM embeddings for species distribution, biodiversity, and habitat classification (Gillespie et al., 2024; Haucke et al., 2025; Huynh et al., 2024; Si-Moussi et al., 2025; Kay et al., 2026). Remote sensing research has long emphasized that the appropriate spatial resolution depends on the structure and scale of the ecological phenomenon being mapped (Woodcock & Strahler, 1987), and that richer spectral and multi-seasonal observations can improve discrimination among vegetation types (Stumpf et al., 2017). At broader spatial scales, combining remotely sensed observations with climatic, topographic, and edaphic information can further improve fine-grained habitat mapping (Si-Moussi et al., 2025; Leblanc et al., 2025). However, systematic understanding of which ecological signals are retained in learned geospatial representations remains limited (Joseph et al., 2026). This is particularly relevant for hierarchical vegetation classification, whose finest levels distinguish communities using floristic composition and local ecological context (FDGC, 2008; Jennings et al., 2009; Sawyer et al., 2009). Such classes exhibit substantial within-class heterogeneity and limited

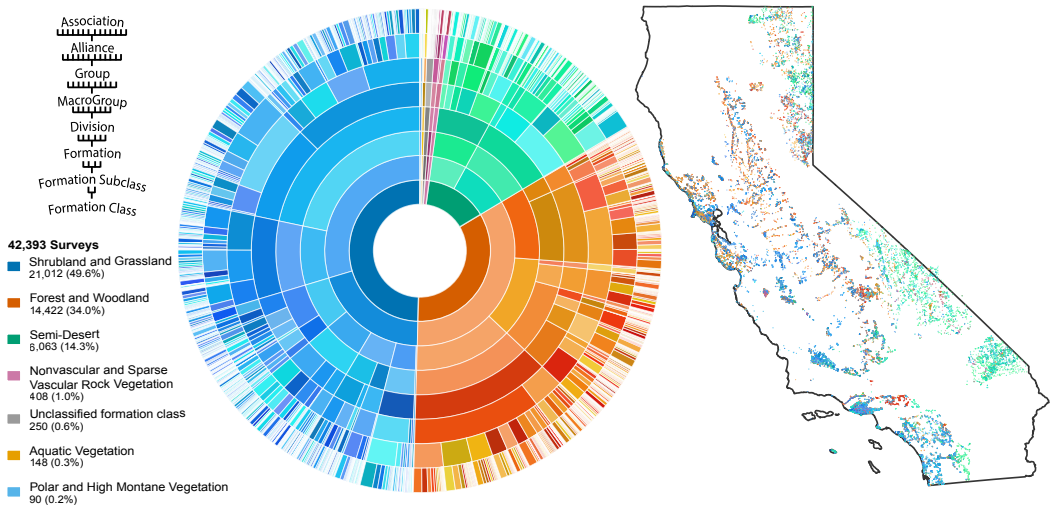

Figure 2: **VEGBENCH leverages vegetation classification and field survey data to evaluate how well geospatial representations capture fine-grained vegetation communities across California.** VEGBENCH uses 42,393 expert field surveys collected across California between 1978 and 2023, labeled at up to eight levels of the U.S. National Vegetation Classification: 6 formation classes, 10 formation subclasses, 17 formations, 22 divisions, 41 macrogroups, 86 groups, 387 alliances, plus 1,598 associations for 29,361 surveys. Survey counts by vegetation category colored iteratively across the vegetation hierarchy are shown in a donut plot on the left; vegetation survey locations colored by group are plotted across California on the right.

spectral separability, making them a challenging yet valuable task (Triepke et al., 2008; Clark et al., 2018; Clark, 2020; Stumpf et al., 2017).

3 THE VEGBENCH BENCHMARK

3.1 BENCHMARK DESIGN

Vegetation hierarchy. Vegetation types in VEGBENCH follow the USNVC v2 hierarchical classification (FDGC, 2008; Franklin et al., 2012), as implemented by VegCAMP and the *Manual of California Vegetation* (Sawyer et al., 2009). This vegetation hierarchy defines eight levels in three broad groups (Fig. 2, examples in Fig. 1). The coarsest three categories are based on dominant growth forms and global macroecological drivers (Formation Class → Formation Subclass → Formation). The next three categories are based on dominant vegetation growth forms and regional meso-ecological drivers (Division → Macrogroup → Group). Finally, the finest categories are based on floristics, diagnostic species, and local environmental conditions (Alliance → Association). In VEGBENCH, we use alliance as the primary vegetation hierarchy prediction target since alliance is the finest level of the hierarchy where the majority of VegCAMP surveys are fully annotated.

Field survey data. In partnership with the California Native Plant Society (CNPS), VEGBENCH is built from vegetation surveys collected by the state of California's VegCAMP program (CDFW, c;a). VegCAMP is an ongoing expert field-survey program that samples vegetation communities across California (CDFW & CNPS, 2025). Surveys are organized into VegCAMP projects, corresponding to vegetation classification and mapping efforts conducted over defined study areas. Vegetation labels and classification practices have changed over the history of VegCAMP. To construct a consistent benchmark, we use *A Manual of California Vegetation Online* (MCV) (CNPS) and the *California Natural Communities List* (CDFW, d) to align historical vegetation survey alliance and association classifications (Jennings et al., 2009) using the most recent publicly available vegetation hierarchy (see Appendix A.2). We additionally crosswalk these labels to the CWHR (Mayer & Laudenslayer, 1988), CALVEG (USDA, Forest Service), and WHR13 habitat-classification systems (CALFIRE, see A.2 for acronym definitions). After label harmonization and quality filtering, we

retain surveys with a valid survey date, a nonempty cleaned alliance label, and spatial coordinates. The resulting benchmark contains 42,393 surveys from 1978 to 2023 across California (Fig. 2).

Benchmark splits. We split vegetation classification surveys within VEGBENCH such that the training, validation, and test sets contain similar proportions of surveys and vegetation alliances from each vegetation mapping project. VegCAMP projects often include accuracy assessment surveys collected after the initial vegetation classification effort, specifically for evaluating draft expert maps (CDFW, 2022). These independently-collected assessments allow for direct comparison to expert-generated vegetation maps, so we additionally reserve all accuracy assessment surveys for testing. To minimize train-test spatial autocorrelation, for the remaining surveys we ensure train and test survey separation by linking observations less than 3.36 km apart so that nearby surveys remain in the same partition. We assign these linked surveys to train, validation, or stratified test, targeting 70%/15%/15% survey proportions overall and within each project, while also preserving the distribution of alliances within projects. Co-registered ablations across stratified versus accuracy assessment surveys showed no strong bias for either validation sampling strategy (see A.6.5, Fig. A5). This results in 23,172 training points, 4,900 validation points, 4,900 stratified test points, and 9,421 accuracy assessment test points (Table A1).

3.2 EVALUATED MODELS

Foundation Models. We evaluate a broad set of foundation models and learned environmental representations, including AlphaEarth (Brown et al., 2025), OlmoEarth v1.3 (Herzog et al., 2026; Team OlmoEarth, 2026), TESSERA v1 (Feng et al., 2026), Prithvi v2 (Szwarcman et al., 2025), CLAY (Clay-Foundation, 2023), SatCLIP (Klemmer et al., 2025a), DINOv3 and DINOv3-Sat (Siméoni et al., 2025), TaxaBind (Sastry et al., 2025), ProM3E (Sastry et al., 2026), CRISP (Huynh et al., 2024), and Deepbiosphere (Gillespie et al., 2024). Model-specific input pre-processing and embedding extraction procedures are described in Appendix A.5.1. These models require diverse remote sensing input sources (Table A2), with NAIP, Sentinel-2, and Landsat source details found in A.4 and A.5.1. Preliminary temporal co-registration ablations (A.3.1) revealed the main disambiguator of model performance was survey count, so for models that require spatiotemporal information we use the exact latitude and longitude from the survey and set a target date of June 15th, 2018.

Baselines. We compare these representations against several standard computer vision baselines, including a trained ResNet-50 (He et al., 2016) and ViT-L/16 (Dosovitskiy et al., 2020) (see Appendix A.5.4), center-pixel probes on Landsat, Sentinel-2, and NAIP data, and MOSAIKS, a fixed random convolutional feature representation derived from NAIP imagery (Rolf et al., 2021). We additionally evaluate with probes fit directly on hand-designed features chosen by experts (see Appendix A.5.2 for additional details), including 21 LiDAR-derived structural features from the NASA WERK program (NASA Earth Exchange (NEX) et al., 2026), 40 different soil properties from POLARIS (Chaney et al., 2019), and simple geographic baselines using latitude/longitude, and latitude/longitude plus elevation (European Space Agency, 2021). Feature extraction, preprocessing, and missing-value handling for hand-designed feature baselines are described in Appendix A.5.2.

California Expert Maps. We also evaluate against vegetation maps produced through VegCAMP's Survey of California Vegetation (SCV) workflow (CDFW, 2022). These maps are created through an expert-in-the-loop process in which vegetation ecologists integrate high resolution aerial imagery (usually NAIP) and environmental data to delineate and attribute vegetation polygons. Initial field vegetation surveys inform the vegetation classification and interpretation key, and determine what alliances are distinguishable—and therefore mappable—from aerial imagery. Mapping from aerial imagery is inherently more difficult and error-prone than ground-level surveys. Therefore, to quantify the error in the mapping process, the draft maps are assessed with the independently-collected, ground-level accuracy assessment surveys. Identified classification errors in draft maps are sometimes corrected prior to finalization and publication, but accuracy changes are often marginal (see A.6.7, Fig. A7).

3.3 TASKS & METRICS

We evaluate each representation for multiclass classification at each level of the vegetation hierarchy, as well as the additional targets CWHR, CALVEG, and WHR13 for direct comparison to common LULC geospatial benchmark tasks (see A.6.10, Fig. A10). For embedding-based models, we fit a

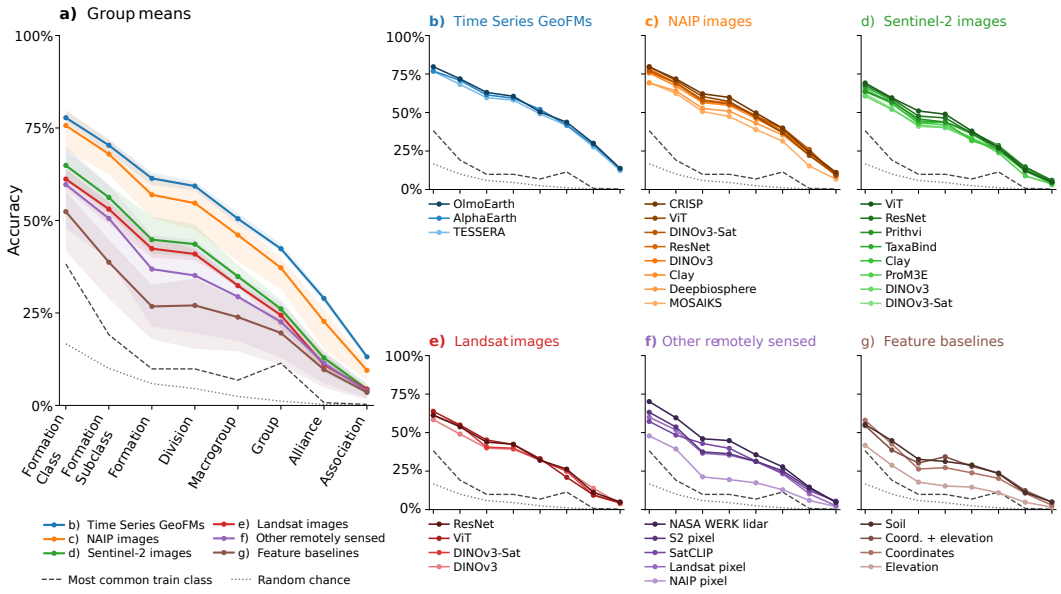

Figure 3: **Performance decreases with increasing ecological granularity. a)** Average accuracy across eight levels of the vegetation hierarchy on the stratified test set by sensor modality, with intervals denoting minimum and maximum performance for models of each category. **b-g)** Individual models within each model type, ordered by mean performance across hierarchy levels.

simple logistic regression to the set of categories at the hierarchy level of interest. We compare with other classification strategies in Appendix Section A.6.6. When evaluating performance we only consider test points with vegetation classes present in the training set (see Table A1). We report:

(1) Accuracy and (2) balanced accuracy, which capture complementary views of classification performance. Accuracy provides the most direct and interpretable measure of how often a model predicts the correct vegetation type. However, because class frequency varies across the vegetation hierarchy, overall accuracy can be disproportionately influenced by more common classes. We therefore also report balanced accuracy, which averages recall equally across classes and provides a complementary measure of whether performance is distributed broadly across vegetation types. When analyzing performance on individual classes, we report per-alliance-class recall.

Lastly, we use **(3) hierarchical accuracy** metric to give partial credit to predictions that are incorrect at the alliance level, but remain correct at a coarser level of the hierarchy. We report hierarchical accuracy after the procedure used by VegCAMP when assessing expert vegetation maps (Menke et al., 2011): predicting a correct alliance at an accuracy assessment point is scored 5/5, correct group 4/5, correct macrogroup 3/5, correct division 2/5, and correct formation 1/5.

4 Results

We group models based on the information available to them: **(1)** geoFMs that utilize time series imagery (OlmoEarth, AlphaEarth, and TESSERA); **(2)** single-image models using high-resolution aerial NAIP imagery; **(3)** single-image Sentinel-2 and **(4)** Landsat models using medium-resolution satellite imagery; **(5)** other remotely sensed baselines, including direct spectral and structural features; and **(6)** feature baselines based on geographic, topographic, and environmental covariates.

How does model performance change across ecological granularity? Across all models and sensor modalities, performance decreases as label resolution goes from coarse formation classes to fine alliances and associations (Fig. 3a). Within sensor groups (Fig. 3b-g), performance clusters tightly, while across sensor groups geoFMs trained with time series data (1) and models using NAIP imagery (2) tend to strongly outperform other approaches. However, as class granularity increases, this gap closes, with Sentinel-2 and Landsat-based embeddings converging to the performance of simple feature baselines. These broad trends also hold for balanced accuracy (Fig. A1), with models per-

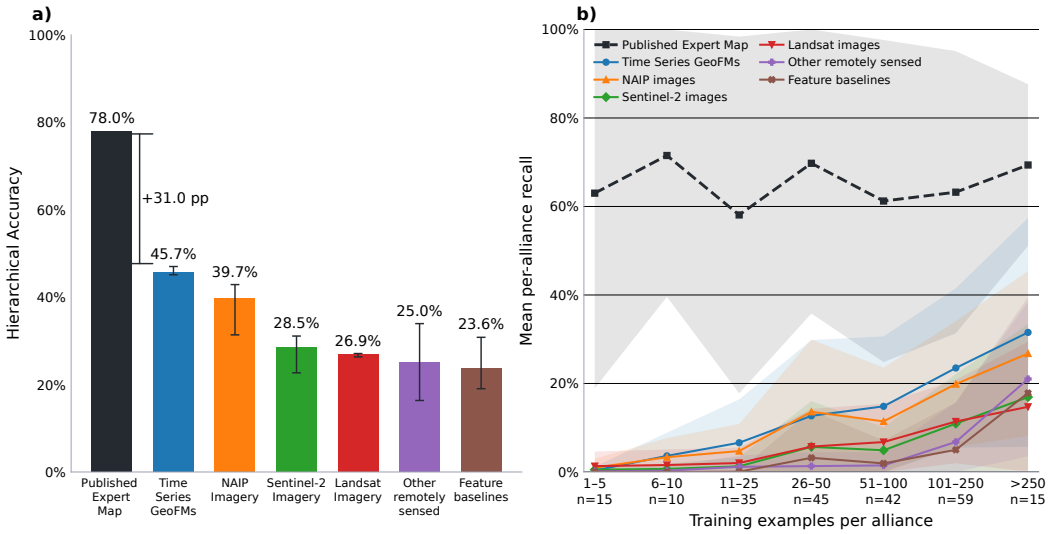

Figure 4: **GeoFM performance is far below expert maps for predicting vegetation alliance. a)** Hierarchical accuracy for accuracy assessment surveys that overlap expert map polygons. Bars are mean across geoFM groupings with whiskers for highest and lowest model performance on the accuracy assessment test set. **b)** Mean per-alliance recall on the accuracy assessment test binned by number of training examples; shaded region indicates standard deviation across binned alliances.

forming better on common and worse on rare classes across the hierarchy. For additional ablations see Appendix Section A.6.

How do geoFMs compare to operational vegetation products? On the independently collected accuracy assessment surveys, published vegetation maps are significantly better than all tested models (+31% published map hierarchical accuracy compared to OlmoEarth, Fig. 4a). We explore this in more detail in the VegCAMP project from Southern Sierra Nevada Foothills (Reyes et al., 2022), which reported both initial map performance on accuracy assessment surveys and performance after experts manually updated these maps post-assessment. This allows us to analyze the performance impact of expert iteration, which improved map accuracy by 5% (Fig. A7). Expert map performance is also consistently high even for rare alliances with little evaluation data (Fig. A8a). In contrast, geoFM performance increases as data availability increases (Fig. 4b), with a linear extrapolation suggesting that time series-based models would need 865-1,062 training surveys per alliance to reach expert map performance (Fig. A8b–c).

What drives geoFM success and failure? Two key traits predominantly drive geoFM performance across VEGBENCH: (1) whether embeddings are an *aggregated time series* and (2) whether models use *high-resolution imagery* when making predictions. Aggregating embeddings across a temporal window appears to broadly improve performance across the vegetation hierarchy regardless of evaluation metric, following (Feng et al., 2026) that aggregation helps stabilize and smooth embeddings across sensor artifacts. Temporal aggregation also appears to help with certain vegetation types where visual appearance can change dramatically across the year, such as deciduous black oaks that distinctly drop their leaves in the winter (Fig. 5ai). These phenological differences are apparent between the winter (leaf-off) and summer (leaf-on) Sentinel-2 observations, but hard to tell apart from high-resolution imagery taken only during leaf-on season when all oak canopies look similar. NAIP image models most often confuse black oak with interior live oak (*Quercus wislizeni–Quercus parvula*; 14% of predictions)—an evergreen oak of the same group that does not drop its leaves in winter—and correctly predict black oak only 16% vs. 45% of the time for time series geoFMs.

However, the large spatial footprint of the Sentinel-2 and Landsat satellite image tiles used by these time series geoFMs can lead to challenges when vegetation alliances are very tightly packed in space, such as river edges where valley oaks often like to grow (Fig. 5biv). Time series geoFMs can struggle when vegetation alliances are mixed within a given satellite tile, while expert mappers excel at delineating these oaks from nearby vegetation (Fig. 5biv, expert map). However, for large pure

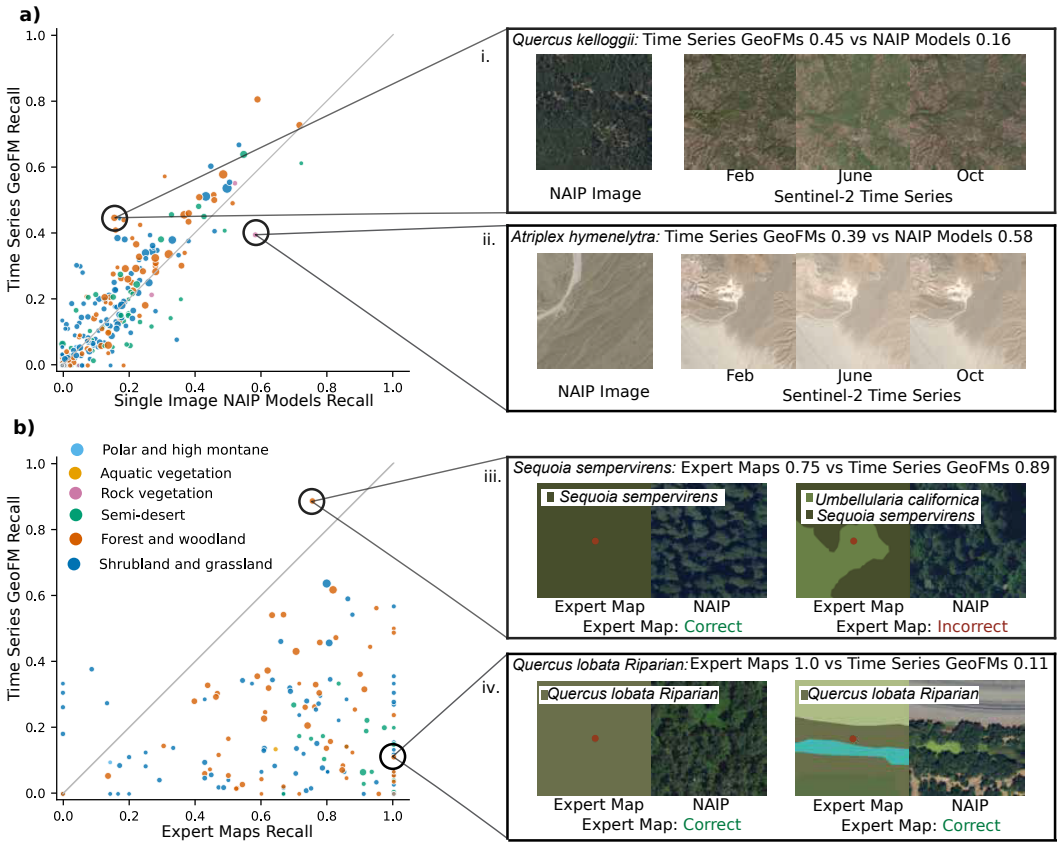

Figure 5: **a) Where Time Series geoFMs and NAIP image models differ most.** The black oak alliance (i) is one example where time series geoFMs outperform NAIP image models, while the desert holly alliance (ii) is one example where NAIP image models outperform time series geoFMs. **b) Where Time Series geoFMs and expert maps differ most.** The redwood alliance (iii) is one example where time series geoFMs outperform expert maps, while the riparian valley oak alliance (iv) is one example where expert maps outperform time series geoFMs. Panel a) pools both test sets (241 alliances with $\geq$5 test surveys); panel b) uses accuracy assessment surveys from vegetation mapping projects with associated expert maps (160 alliances).

vegetation stands expert maps can sometimes overestimate boundaries between vegetation classes, leading to better performance by geoFMs for large alliances like redwood forest (Fig. 5biii). Interestingly, vegetation stand purity within a given radius of a survey appears to be less important for OlmoEarth and supervised approaches trained with NAIP and Landsat imagery, while strongly influencing other models (A.6.3, Fig. A3).

In contrast, access to high-resolution imagery exclusively improves performance over medium-resolution imagery models (A.6.2, Fig. A2). However, in comparison to temporal embedding aggregation, high-resolution imagery improves performance sometimes, but not always (Fig. 5a). For certain vegetation types with distinctive but very small lifeforms like desert holly (Fig. 5ai), the high spatial resolution of NAIP imagery is uniquely able to capture signatures of the bushes that are too fine-grained for Sentinel-2 imagery to resolve. When controlling for model architecture and training strategy, the performance advantage of higher spatial resolution NAIP data is even more apparent, with performance often dropping by half or more when going from 0.6 m-resolution NAIP data to 10m- and 30m-resolution Sentinel-2 and Landsat data (Fig. A2). The only outlier is the simple baseline of training on raw reflectance values at the pixel centered over a survey, where the few number of bands and high spatial resolution of NAIP imagery becomes a disadvantage.

5 DISCUSSION

What current geoFMs capture. Across every model family, vegetation classification performance falls steadily as labels move from broad formation classes to alliances and associations (Fig. 3). Strong accuracy on coarse ecological categories therefore does not imply that a representation encodes the structural distinctions necessary to define vegetation communities. At the finest levels, embeddings from many models—especially those generated with medium-resolution imagery data—perform no better than simple geographic and environmental baselines, suggesting that much of the signal they capture is currently location rather than vegetation. Evaluating only coarse biome or land-cover classification would miss this gap entirely, and most existing geospatial benchmarks do.

Where gaps for geoFMs remain. Aggregating imagery through time and using sub-meter imagery appears to improve vegetation classification performance in complementary ways, but no prominent geoFM combines both. Time series embeddings can recover vegetation easily distinguished by phenology, such as deciduous oak woodland, while high-resolution images disambiguate vegetation defined by fine structure, such as sparse desert shrubs (Fig. 5). In contrast, access to sub-meter imagery significantly improves performance in essentially all cases when controlling for architecture and training strategy (Fig. A2). Notably, NAIP is the same imagery expert mappers interpret to build their maps, indicating that the information needed for fine-grained mapping is often present in NAIP imagery but not fully leveraged by current models. Rare vegetation alliances also remain an open problem: below ~ 10 training examples even the strongest representations approach zero accuracy per-alliance (Fig. 4b), and no model family consistently succeeds on these alliances. Unlike geoFMs, expert maps can manage to reach high accuracy even for some of the rarest vegetation alliances (Fig. A.6.8a). Additionally, at high ecological granularity, LiDAR-based baselines (NASA WERK LiDAR, Fig. 3) perform competitively with many more complex models, suggesting that the structural information LiDAR provides is helpful but underutilized by current geoFMs.

What operational vegetation mapping suggests. Currently, expert vegetation maps are far more accurate than any geospatial model, and their accuracy is often independent of an alliance's frequency in the training data (Fig. 4b). In contrast, geoFM accuracy improves substantially with the number of labeled examples. Extrapolating this trend linearly, geoFMs would need nearly 1,000 surveys per-alliance to match expert maps (Fig. A8b–c), several times more than even the most common alliances have today. Also, at roughly $1,000 per survey (Ratchford et al., 2026), closing the gap through data collection alone would cost upwards of USD$300 million. The high accuracy of expert vegetation maps—especially on rare classes— suggests that mirroring expert vegetation mapmaking procedures in geoFM training could be a fruitful research direction for the community. These procedures include combining high-resolution imagery with field knowledge, ground level images, classification keys, and ancillary environmental data, plus assigning a coarser category when an alliance cannot be resolved. Representations that fuse spatial and temporal signals, incorporate ground-level imagery, leverage natural text descriptions, or better represent uncertainty are promising directions for methodological innovation that VEGBENCH can evaluate.

Expanding VEGBENCH to broader geographies. The vegetation classification and field surveys underpinning VEGBENCH have been collected in places well beyond California, including all of Europe (Chytrý et al., 2020) and South Africa (Dayaram et al., 2019). Our benchmark provides a scalable path toward expanding VEGBENCH globally to incorporate additional geographies and ecological communities through this global, unified vegetation hierarchy (Keith et al., 2020).

Limitations. VEGBENCH is restricted to California and to a single—though widely adopted—hierarchical vegetation classification, so absolute performance may not transfer to other regions or classification systems. The field surveys were collected to build and validate vegetation maps rather than to evaluate learned representations, and the underlying vegetation projects differ in their sampling design, era, and mapping methods. Survey locations are also not a uniformly random sample of California. We also only evaluate frozen embeddings with no fine-tuning considered.

Conclusion. VEGBENCH evaluates geospatial representations across an ecologically-relevant hierarchy using expert field surveys. Current representations capture broad ecological structure but fall far short of the fine-grained distinctions required for operational vegetation mapping, with large gaps remaining for rare, structurally, and phenologically distinctive communities. By making ecological resolution measurable, VEGBENCH provides a testbed for developing geospatial models that can support the vegetation mapping conservationists and land managers need.

AI USE STATEMENT

In this work, we used generative AI tools for:

- Implementing the research methods.
- Cleaning datasets or changing their format.
- Analyzing qualitative data or identifying themes.
- Interpreting research findings.

We have not used generative AI tools for

- Developing theoretical models or conceptual frameworks.
- Developing or refining research hypotheses.
- Designing research methods or experiments, or giving feedback on them.

Producing synthetic datasets, formulating mathematical statements or conjectures, supplying essential steps or ideas for mathematical proofs, writing proofs, and translating material are not applicable to this work.

Additionally, we used generative AI tools for:

- Producing or modifying scientific illustrations.
- Selecting experiment parameters.
- Writing or editing software.
- Producing research artifacts.
- Searching for information.
- Improving manuscript readability.
- Finding relevant publications.
- Formatting citations and bibliographies.
- Planning the manuscript's organization.

We have reviewed all AI-assisted work. We manually checked all LLM-formatted citations, verified and tested for LLM-generated code correctness, manually verified scientific illustrations for correctness and scientific accuracy, swept LLM-seeded hyperparameters, and manually verified dataset cleaning pipelines. The final manuscript has been fully read by multiple authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All results presented in this manuscript are fully reproducible from the code and dataset submitted in the supplemental materials. All code and primary artifacts to generate the benchmark dataset are in the supplemental materials. Remote sensing download, embedding generation, and probe fitting has been verified to work on two separate machines. All figures and tables are sufficiently described to be reproducible using the benchmark pipeline. All code as well as the benchmark dataset will be released publicly upon publication.

REFERENCES

Kenneth A Baldwin, K Chapman, D Meidinger, P Uhlig, L Allen, S Basquill, D Faber-Langendoen, N Flynn, C Kennedy, W Mackenzie, et al. *The Canadian national vegetation classification: principles, methods and status*. Natural Resources Canada, Canadian Forest Service, Great Lakes Forestry Centre, 2019.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Jesslyn F Brown, Heather J Tollerud, Christopher P Barber, Qiang Zhou, John L Dwyer, James E Vogelmann, Thomas R Loveland, Curtis E Woodcock, Stephen V Stehman, Zhe Zhu, et al. Lessons learned implementing an operational continuous united states national land change monitoring capability: The land change monitoring, assessment, and projection (lcmap) approach. *Remote sensing of environment*, 238:111356, 2020.

CALFIRE. California vegetation - WHR13 types. Geospatial dataset. URL https://data.ca.gov/dataset/california-vegetation-whr13-types. Wildlife Habitat Relationship classes grouped into 13 major land cover types; accessed 2026-05-14.

CDFW. Vegetation survey points [ds1020]. GIS dataset, a. URL https://filelib.wildlife.ca.gov/public/BDB/GIS/BIOS/metadata/DS1020.html. Biogeographic Information and Observation System (BIOS). Accessed 2026-05-14.

CDFW. California wildlife habitat relationships. California Wildlife Habitat Relationships System, b. URL https://wildlife.ca.gov/Data/CWHR/Data/CWHR/Wildlife-Habitats. Version 10. Accessed 2026-05-14.

CDFW. Vegetation classification and mapping program (VegCAMP). Biogeographic Data Branch, California Department of Fish and Wildlife, c. URL https://wildlife.ca.gov/Data/VegCAMP. Accessed 2026-05-14.

CDFW. California vegetation classification: Alliance and association names. Microsoft Excel spreadsheet, d. URL https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=153399. Accessed 2026-05-14.

CDFW. VegCAMP background. https://wildlife.ca.gov/Data/VegCAMP/Background, e. Accessed: 2026-08-15.

CDFW. A status review of the fisher (Martes pennanti) in california. Report to the fish and game commission, California Department of Fish and Wildlife, February 2010. URL https://www.nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=83920.

CDFW. Survey of california vegetation: Classification and mapping standards. Technical report, California Department of Fish and Wildlife, November 2022. URL https://filelib.wildlife.ca.gov/Public/BDB/Vegetation/ClassificationAnalysisRFP/Attachment1_SCV%20Classification%20and%20Mapping%20Standards%2020221104.pdf. Version dated November 4, 2022.

CDFW. California state wildlife action plan, swap 2025: A conservation legacy for californians. State wildlife action plan, California Department of Fish and Wildlife, 2025. URL https://wildlife.ca.gov/SWAP. Accessed 2025-12-15.

CDFW and CNPS. A shared vision for the survey of california vegetation. Report, California Department of Fish and Wildlife, Sacramento, CA, 2014. URL https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=114778. Available via the VegCAMP framework.

CDFW and CNPS. CDFW-CNPS protocol for the combined vegetation rapid assessment and relevé field form. Technical report, California Department of Fish and Wildlife and California Native Plant Society, May 2025. URL https://www.nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=18599. Version dated May 16, 2025.

Nathaniel W. Chaney, Budiman Minasny, Jonathan D. Herman, Travis W. Nauman, Colby W. Brungard, Cristine L. S. Morgan, Alexander B. McBratney, Eric F. Wood, and Yohannes Yimam. POLARIS properties: 30-meter probabilistic maps of soil properties over the contiguous united states. *Water Resources Research*, 2019. doi: 10.1029/2018WR022797. URL https://doi.org/10.1029/2018WR022797.

Milan Chytrỳ, Lubomír Tichỳ, Stephan M Hennekens, Ilona Knollová, John AM Janssen, John S Rodwell, Tomáš Peterka, Corrado Marcenò, Flavia Landucci, Jiří Danihelka, et al. Eunis habitat classification: Expert system, characteristic species combinations and distribution maps of european habitats. *Applied Vegetation Science*, 23(4):648–675, 2020.

Matthew L Clark. Comparison of multi-seasonal landsat 8, sentinel-2 and hyperspectral images for mapping forest alliances in northern california. *ISPRS journal of photogrammetry and remote sensing*, 159:26–40, 2020.

Matthew L Clark, Jennifer Buck-Diaz, and Julie Evens. Mapping of forest alliances with simulated multi-seasonal hyperspectral satellite imagery. *Remote sensing of environment*, 210:490–507, 2018.

Clay-Foundation. Clay, 2023. URL https://doi.org/10.57967/hf/1537.

Frederic Edward Clements. *Plant indicators: the relation of plant communities to process and practice*. Number 290. Carnegie Institution of Washington, 1920.

CNPS. A manual of california vegetation online. Online vegetation classification database. URL https://vegetation.cnps.org/classifications. Accessed 2026-07-02.

Anisha Dayaram, Linda R Harris, Barend A Grobler, Stephni Van der Merwe, Anthony G Rebelo, Leslie Ward Powrie, Johannes HJ Vlok, Philip G Desmet, Mcebisi Qabaqaba, Keneilwe M Hlahane, et al. Vegetation map of south africa, lesotho and swaziland 2018: a description of changes since 2006. *Bothalia-African Biodiversity & Conservation*, 49(1):1–11, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Raphaël d'Andrimont, Momchil Yordanov, Laura Martinez-Sanchez, Beatrice Eiselt, Alessandra Palmieri, Paolo Dominici, Javier Gallego, Hannes Isaak Reuter, Christian Joebges, Guido Lemoine, et al. Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union. *Scientific data*, 7(1):352, 2020.

EEA. EUNIS Terrestrial Habitat Classification Review. Version 1, November 2021, 2021. URL https://doi.org/10.2909/bfe4c237-e378-4a83-ab21-b3807f96c2e2. Updated January 2023.

European Space Agency. Copernicus harmonized sentinel-2 msi: Multispectral instrument, level-2a (sr).

European Space Agency. Copernicus Digital Elevation Model (GLO-30), 2021. URL https://doi.org/10.5270/ESA-c5d3d65. Copernicus DEM GLO-30 global digital surface model, 30 m resolution.

Don Faber-Langendoen, Todd Keeler-Wolf, Del Meidinger, Dave Tart, Bruce Hoagland, Carmen Josse, Gonzalo Navarro, Serguei Ponomarenko, Jean-Pierre Saucier, Alan Weakley, et al. Ecoveg: a new approach to vegetation description and classification. *Ecological Monographs*, 84(4):533–561, 2014.

Don Faber-Langendoen, Ken Baldwin, Robert K Peet, Del Meidinger, Este Muldavin, Todd Keeler-Wolf, Carmen Josse, et al. The ecoveg approach in the americas: Us, canadian and international vegetation classifications. *Phytocoenologia*, 48(2):215–237, 2018.

Don Faber-Langendoen, David A Keith, Javier Loidi, Eileen H Helmer, Wolfgang Willner, Gonzalo Navarro, John Hunter, Changcheng Liu, Reginald T Guuroh, and Patricio Pliscoff. Advancing the ecoveg approach as a terrestrial ecosystem typology: From global biomes to local plant communities. *Ecosphere*, 16(5):e70237, 2025.

FDGC. National vegetation classification standard, version 2.0. Technical Report FGDC-STD-005-2008, Federal Geographic Data Committee, Reston, Virginia, 2008. URL https://www.fgdc.gov/standards/projects/vegetation/NVCS_V2_FINAL_2008-02.pdf.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026.

Mark A Finney. Farsite: Fire area simulator-model development and evaluation. *Res. Pap. RMRS-RP-4, Revised 2004. Ogden, UT: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 47 p.*, 4, 1998.

SB Franklin, D Faber-Langendoen, M Jennings, T Keeler-Wolf, O Loucks, R Peet, D Roberts, and A McKerrow. Building the united states national vegetation classification. *Annali di Botanica*, 2: 1–9, 2012.

Lauren E Gillespie, Megan Ruffley, and Moises Exposito-Alonso. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, 121(37):e2318296121, 2024.

Timm Haucke, Lauren Harrell, Yunyi Shen, Levente Klein, David Rolnick, Lauren Gillespie, and Sara Beery. Deep multi-modal species occupancy modeling. *bioRxiv*, pp. 2025–09, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34806–34817, 2026.

Andy V Huynh, Lauren E Gillespie, Jael Lopez-Saucedo, Claire Tang, Rohan Sikand, and Moisés Expósito-Alonso. Contrastive ground-level image and remote sensing pre-training improves representation learning for natural world imagery. In *European Conference on Computer Vision*, pp. 173–190. Springer, 2024.

Takayuki Ishikawa, Carmelo Bonannella, Bas JW Lerink, and Marc Rußwurm. Assessing the effectiveness of deep embeddings for tree species classification in the dutch forest inventory. *Science of Remote Sensing*, pp. 100498, 2026.

Michael D. Jennings, Don Faber-Langendoen, Orie L. Loucks, Robert K. Peet, and David Roberts. Standards for associations and alliances of the U.S. National Vegetation Classification. *Ecological Monographs*, 79(2):173–199, 2009. ISSN 1557-7015. doi: 10.1890/07-1804.1.

Maxwell B Joseph, Flávia De Souza Mendes, Dieu My T Nguyen, Camile Sothe, and Christopher B Anderson. Continuous biome representations from earth observation embeddings. *arXiv preprint arXiv:2606.11510*, 2026.

Justin Kay, Shir Bar, Ellen O. Aikens, Martin Becker, Francesca Cagnacci, Juliet Cohen, Scott W. Forrest, Jessica Kendall-Bar, Madeleine Lucas, Macon Overcast, Meredith S. Palmer, Will Rogers, Nicholas J. Russo, Christian Rutz, Larissa T. Beumer, Michael Brown, Ying-Chi Chan, Sarah C. Davidson, Diego Ellis Soto, Anne G. Hertel, Roland Kays, Benjamin Koger, Guram Mikaberidze, Thomas Mueller, Ruth Oliver, Thorsten Papenbrock, Robert Patchett, Jared A. Stabach, Dane Taylor, Scott W. Yanco, and Sara Beery. Movebench: A benchmark for global-scale wildlife movement forecasting, 2026. URL `https://arxiv.org/abs/2609.15780`.

David A. Keith, Jose R. Ferrer-Paris, Emily Nicholson, and Richard T. Kingsford (eds.). *IUCN Global Ecosystem Typology 2.0: Descriptive Profiles for Biomes and Ecosystem Functional Groups*. IUCN, International Union for Conservation of Nature, December 2020. ISBN 978-2-8317-2077-7. doi: 10.2305/IUCN.CH.2020.13.en.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4347–4355, 2025a.

Konstantin Klemmer, Esther Rolf, Marc Russwurm, Gustau Camps-Valls, Mikolaj Czerkawski, Stefano Ermon, Alistair Francis, Nathan Jacobs, Hannah Rae Kerner, Lester Mackey, Gengchen Mai, Oisin Mac Aodha, Markus Reichstein, Caleb Robinson, David Rolnick, Evan Shelhamer, Vincent Sitzmann, Devis Tuia, and Xiaoxiang Zhu. Earth Embeddings: Towards AI-centric Representations of our Planet. December 2025b.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

César Leblanc, Lukas Picek, Rémi Palard, Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, and Alexis Joly. Mapping biodiversity at very-high resolution in europe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2374–2383, 2025.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models, April 2025.

Kenneth E Mayer and William F Laudenslayer. *A guide to wildlife habitats of California*. The Department, 1988.

John Menke, Ed Reyes, Debbie Johnson, Julie Evens, and Kendra Sikes. Northern Sierra Nevada Foothills Vegetation Project: Vegetation Mapping Report. Technical report, CDFW, February 2011.

Ladislav Mucina. Classification of vegetation: Past, present and future. *Journal of Vegetation Science*, 8(6):751–760, 1997.

NASA Earth Exchange (NEX), California Natural Resources Agency (CNRA), and California Air Resources Board (CARB). NASA WERK LiDAR: California Statewide (30 m), 2026. URL https://wifire-data.sdsc.edu/dataset/nasa-werk-dem-statewide-30m. Wildfire, Ecosystem Resilience, and Risk Assessment (WERK) Initiative. LiDAR-derived statewide topography and ecosystem structure products.

David M Olson, Eric Dinerstein, Eric D Wikramanayake, Neil D Burgess, George VN Powell, Emma C Underwood, Jennifer A D'amico, Illanga Itoua, Holly E Strand, John C Morrison, et al. Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938, 2001.

James M Omernik and Glenn E Griffith. Ecoregions of the conterminous united states: evolution of a hierarchical spatial framework. *Environmental management*, 54(6):1249–1266, 2014.

Arjun Rao, Marc Rußwurm, Konstantin Klemmer, and Esther Rolf. Measuring the intrinsic dimension of earth representations. In *International Conference on Learning Representations*, volume 2026, pp. 113200–113224, 2026.

Jaime Ratchford, Rachelle Boul, Betsy Harbert, and Rosalie Yacoub. A historic investment in the survey of california vegetation and keeping the momentum. Talk 18.2 presented in the session "30 Years of *A Manual of California Vegetation*," CNPS Conference, February 2026. URL https://conference.cnps.org/sessions-old/18-30-years-of-a-manual-of-california-vegetation/. Friday, February 6; Hall B; session time 1:00–2:45 pm.

Edward Reyes, Arin Glass, Julie Evens, Jennifer Buck-Diaz, Alexis LaFever-Jackson, Jaime Ratchford, Rachelle Boul, Rosalie Yacoub, John Menke, John Fulton, and Deborah Johnson. Fine-scale vegetation map and accuracy assessment of the Southern Sierra Nevada Foothills, California. Final Report Contract P1884008, Aerial Information Systems, Inc., Redlands, CA, March 2022. URL https://nrm.dfg.ca.gov/FileHandler.ashx?DocumentID=211115. Prepared for the California Department of Fish and Wildlife.

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(1):4392, 2021.

Matthew G Rollins. Landfire: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, 18(3):235–249, 2009.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1765–1774. IEEE, 2025.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Jiayu Lin, Dan Cher, Phoenix Jarosz, and Nathan Jacobs. Prom3e: Probabilistic masked multimodal embedding model for ecology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20564–20574, 2026.

John O. Sawyer, Todd Keeler-Wolf, and Julie M. Evens. *A Manual of California Vegetation*. California Native Plant Society, Sacramento, California, 2 edition, 2009. ISBN 978-0-943460-49-9.

Sara Si-Moussi, Stephan Hennekens, Sander Mücher, Wanda De Keersmaecker, Milan Chytrý, Emiliano Agrillo, Fabio Attorre, Idoia Biurrun, Gianmaria Bonari, Andraž Čarni, Renata Ćušterevska, Tetiana Dziuba, Klaus Ecker, Behlül Güler, Ute Jandt, Borja Jiménez-Alfaro, Jonathan Lenoir, Jens-Christian Svenning, Grzegorz Swacha, and Wilfried Thuiller. EUNIS habitat maps: Enhancing thematic and spatial resolution for Europe through machine learning. *Scientific Data*, 12(1):1976, December 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-06235-7.

Sara Si-Moussi, Stephan Hennekens, Sander Mucher, Stan Los, Yoann Cartier, Borja Jiménez-Alfaro, Fabio Attorre, Jens-Christian Svenning, and Wilfried Thuiller. Continental-scale habitat distribution modelling with multimodal earth observation foundation models. *arXiv preprint arXiv:2507.09732*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longepe, Xiao Xiang Zhu, Hannah Kerner, Juan Bernabe-Moreno, and Alexandre Lacoste. GEO-Bench-2: From Performance to Capability, Rethinking Evaluation in Geospatial AI. (arXiv:2511.15658), February 2026. doi: 10.48550/arXiv.2511.15658.

Jeffrey R. Smith, Andrew D. Letten, Po Ju Ke, Christopher B. Anderson, J. Nicholas Hendershot, Manpreet K. Dhami, Glade A. Dlott, Tess N. Grainger, Meghan E. Howard, Beth M.L. Morrison, Devin Routh, Priscilla A. San Juan, Harold A. Mooney, Erin A. Mordecai, Thomas W. Crowther, and Gretchen C. Daily. A global test of ecoregions. *Nature Ecology and Evolution*, 2(12):1889–1896, 2018. ISSN 2397334X. doi: 10.1038/s41559-018-0709-x.

Radost Stanimirova, Katelyn Tarrio, Konrad Turlej, Kristina McAvoy, Sophia Stonebrook, Kai-Ting Hu, Paulo Arévalo, Eric L Bullock, Yingtong Zhang, Curtis E Woodcock, et al. A global land cover training dataset from 1984 to 2020. *Scientific Data*, 10(1):879, 2023.

Kenneth A. Stumpf, Cogan Technology, and Kier Associates. Vegetation mapping and classification project: Redwood national and state parks, california. Natural Resource Report NPS/REDW/NRR–2017/1431, National Park Service, Fort Collins, Colorado, May 2017.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Team OlmoEarth. OlmoEarth v1.3: Embeddings. Technical report, Allen Institute for AI, 2026. URL `https://olmoearth.allenai.org`.

Wilfried Thuiller, David M Richardson, Petr Pyšek, Guy F Midgley, Greg O Hughes, and Mathieu Rouget. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global change biology*, 11(12):2234–2250, 2005.

F Jack Triepke, C Kenneth Brewer, Daniel M Leavell, and Stephen J Novak. Mapping forest alliances and associations using fuzzy systems and nearest neighbor classifiers. *Remote Sensing of Environment*, 112(3):1037–1050, 2008.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

U.S. Department of Agriculture, Farm Service Agency. National agriculture imagery program (naip).

U.S. Geological Survey. Landsat collection 2 level-2.

USDA, Forest Service. CALVEG: Classification and assessment with LANDSAT of visible ecological groupings. Existing vegetation classification and mapping system. URL `https://data.fs.usda.gov/geodata/edw/datasets.php?xmlKeyword=calveg`. Accessed 2026-05-14.

USDA-NASS. Cropland data layer, 2025. URL `https://croplandcros.scinet.usda.gov/`. USDA NASS Cropland Data Layer (CDL).

USNVC. United States National Vegetation Classification Database, Version 3.0, 2025. URL `https://usnvc.org/usnvc-3-0-catalog/`. USNVC Hierarchy Explorer; accessed 25 September 2026.

Curtis E. Woodcock and Alan H. Strahler. The factor of scale in remote sensing. *Remote Sensing of Environment*, 21(3):311–332, April 1987. ISSN 00344257. doi: 10.1016/0034-4257(87)90015-0.

A APPENDIX

A.1 TERMS AND ACRONYMS

Vegetation and survey terms.

- **Vegetation alliance:** A recurring assemblage of plant species that co-occur under similar environmental conditions.
- **Vegetation classification:** A hierarchical system that names vegetation communities, from broad structural types down to alliances defined by their species.
- **Physiognomy (physiognomic):** The physical structure of vegetation, such as its dominant growth form (tree, shrub, herb) and height, rather than which species are present. The upper levels of the hierarchy are physiognomic.
- **Floristic:** Relating to the plant species that make up a community. The lower levels of the hierarchy are floristic.
- **Diagnostic species:** The species whose presence, dominance, or combination distinguishes one vegetation class from similar ones.
- **Formation class, subclass, and formation:** The three upper vegetation hierarchy levels, distinguished by dominant growth form and broad climatic and ecological conditions (e.g., forest and woodland vs. shrubland and grassland).
- **Division, macrogroup, and group:** The three middle vegetation hierarchy levels, which add regional biogeography and increasingly specific sets of characteristic species.
- **Alliance:** A fine floristic level defined by one or more diagnostic species, usually the dominant species of the uppermost vegetation layer, together with its typical habitat. Alliances are the main fine-grained target in VEGBENCH.
- **Association:** The finest level, which subdivides an alliance by the species of its lower layers (e.g., shrub or herb understory) and the local site conditions they reflect.
- **Example path through the hierarchy:** A blue oak woodland survey is labeled Mesomorphic Tree Vegetation (formation class) → Temperate Forest (subclass) → Warm Temperate Forest (formation) → Madrean Forest and Woodland (division) → California Forest and Woodland (macrogroup) → Californian broadleaf forest and woodland (group) → *Quercus douglasii* (alliance, defined by the dominant tree, blue oak) → *Quercus douglasii* / *Bromus* spp.–(*Daucus pusillus*) (association, a blue oak canopy over annual grasses). Other associations in the same alliance have the same canopy over a different understory, such as the shrub *Ceanothus cuneatus*.
- **Field survey:** An observation made on the ground at a surveyed location, where trained ecologists record the species present and their cover and assign a vegetation class.
- **Relevé and rapid assessment:** The two main VegCAMP survey protocols. A relevé records all plant species and their cover within a fixed plot; a rapid assessment records the dominant and characteristic species more quickly with ocular estimates and is used to cover more sites.
- **accuracy assessment (AA) survey:** A field survey collected after a draft vegetation map exists, specifically to measure the map's accuracy. The field crew is not shown the mapped label, so these surveys are independent of the map.
- **Expert vegetation map:** A map that divides a study area into polygons and assigns each polygon a vegetation class. Ecologists draw it by photointerpretation of aerial imagery, guided by field surveys and environmental data, or use classical supervised machine learning approaches.
- **photointerpretation:** Delineating and labeling vegetation by visually interpreting aerial or satellite imagery.
- **Crosswalk:** A mapping between classes of two classification systems, e.g. going from vegetation alliance to CWHR, CALVEG, or WHR13 habitat types.
- **Land cover:** A broad description of what covers the surface (e.g., forest, grassland, water, urban). Land-cover classes are much coarser than vegetation communities.
- **Ecoregion and biome:** Large geographic regions with broadly similar climate, landforms, and biota (e.g., Olson ecoregions); each typically contains many vegetation communities.
- **Phenology:** Seasonal change in vegetation, such as leaf-out, flowering, or leaf drop. Deciduous plants shed their leaves seasonally; evergreen plants keep them year-round.

- **Hierarchical accuracy:** A partial-credit score in which a wrong alliance prediction still earns credit if it belongs to the correct group, macrogroup, division, or formation (Section 3.3).

Acronyms.

- **AA:** Accuracy assessment
- **CALFIRE:** California Department of Forestry and Fire Protection
- **CALVEG:** Classification and Assessment with Landsat of Visible Ecological Groupings
- **CDFW:** California Department of Fish and Wildlife
- **CNPS:** California Native Plant Society
- **CWHR:** California Wildlife Habitat Relationships
- **EEA**: European Environment Agency
- **EPA:** Environmental Protection Agency
- **EUNIS:** European Nature Information System
- **FGDC:** Federal Geographic Data Committee
- **GEE:** Google Earth Engine
- **GeoFM:** Geospatial foundation model
- **IUCN:** International Union for Conservation of Nature
- **LANDFIRE:** Landscape Fire and Resource Management Program
- **MCV:** *A Manual of California Vegetation*
- **NAIP:** National Agriculture Imagery Program (0.6 m aerial imagery)
- **SCV:** Survey of California Vegetation
- **USDA-NASS:** U.S. Department of Agriculture National Agricultural Statistics Service
- **USNVC:** U.S. National Vegetation Classification
- **VegCAMP:** Vegetation Classification and Mapping Program
- **WHR13:** Wildlife Habitat Relationships 13 major land-cover types

A.2 Survey Data Cleaning and Label Harmonization

We standardized the vegetation survey data from the VegCAMP DS1020 geodatabase (CDFW, a) prior to constructing benchmark splits. The source data contained vegetation labels collected across multiple projects and classification systems, with variation in field names, taxonomic level, and nomenclature. We therefore developed a cleaning pipeline to harmonize these observations against California Native Plant Society (CNPS) manual of California vegetation (Sawyer et al., 2009) while retaining the original recorded alliance name for provenance.

We first standardized heterogeneous source fields to a common schema, including MCV and NVCS vegetation names and hierarchy fields, California alliance codes, CALVEG (USDA, Forest Service) classes, and CWHR (CDFW, b) habitat classes. String values were normalized for whitespace, punctuation, and common classification terms prior to matching. Missing or non-informative values such as “N/A”, “unknown”, “not mapped”, and “not assigned” were treated as missing.

To establish a canonical vegetation hierarchy, we constructed a CNPS crosswalk from CNPS alliance records and CNPS alliance pages. The crosswalk linked modern alliance names and California alliance codes to historical MCV (Sawyer et al., 2009) and NVCS (Franklin et al., 2012) nomenclature and to the corresponding group, macrogroup, division, formation, formation subclass, and formation class. It additionally retained CNPS links to CalVeg and CWHR habitat classes where available.

For each survey point, alliance labels were resolved using a hierarchical matching procedure. We first attempted exact matches using available MCV and NVCS alliance or vegetation-name fields and California vegetation codes. A match was accepted only when it identified a unique alliance. For labels without an exact reference match, we allowed approximate name matching with a minimum string-similarity threshold of 0.9, again requiring the resulting candidate set to identify a unique alliance. Remaining labels could be resolved against CNPS alliance-page information using agreement with the source vegetation label.

Association-level observations were handled separately. When a survey was explicitly labeled at the association level, association names and codes were matched against the CNPS association reference (CNPS). Cleaned association labels were populated only for reference matches; matched associations could also supply the parent alliance hierarchy.

CalVeg and CWHR labels were standardized against their respective reference tables, prioritizing recognized codes over normalized class names. CWHR standardization additionally corrected predefined code and habitat-name aliases and recovered codes from recognized habitat names when source codes were missing or unrecognized. When both the cleaned code and class name remained empty, corresponding classifications from the CNPS page for the resolved alliance were used where available. Multiple classifications were retained.

We additionally corrected clearly implausible surveys. Survey dates with years prior to 1975 were replaced only when a unique modal year could be identified from other observations in the same project, preferentially using records with the same month and day and otherwise using the project-wide modal year. Dates for which no unique replacement year could be established were left unchanged. We also removed all observations in groups sharing identical point coordinates and survey timestamps when they resolved to conflicting non-empty alliance labels.

Each source feature was assigned a persistent survey-point identifier. Point geometry was transformed to California Albers (EPSG:3310), while geographic longitude and latitude coordinates in WGS84 were also retained for subsequent remote-sensing data retrieval.

A.3 Additional Benchmark Details

A.3.1 Temporal Alignment

Given the large temporal extent of the vegetation surveys with the relatively recent availability of remote sensing products, including Sentinel-2 availability beginning in 2017, we consider three temporal-alignment settings for training:

Co-registered subset: We only consider surveys taken 2017 or later and temporally align each model's input to the survey to the extent supported by the model, resulting in only 3,996 training surveys covering 278 unique alliances and 681 unique associations.

Fixed date: We use all surveys but pull model inputs at a common reference date of June 15, 2018. This setting sacrifices survey-image temporal correspondence in exchange for a substantially larger labeled training set: all 23,172 training surveys covering 379 unique alliances and 1,485 unique associations.

Fixed date subset: We use only surveys collected in 2017 or later but pull model inputs at the same fixed reference date of June 15, 2018. This allows us to disentangle the benefits of training data size with temporal co-registration.

These temporal alignment strategies can also impact the test and validation sets. When comparing across training settings, we use the subset of test surveys collected in 2017 or later with the same model input source as the training set. In the main text, we report results on the **fixed date** temporal-alignment setting because it consistently produced the best results; see A.6.4 and Fig. A4 for results and further details.

A.3.2 Data Splitting

Training uses 23,172 surveys (3,996 from 2017 or later) and validation 4,900 (447 from 2017 or later). Because we evaluate supervised models, test surveys whose class never appears in training are excluded from scoring: at the alliance level only 4 stratified and 6 accuracy assessment surveys, but at the association level 219 and 92. These unseen classes, together with the many alliances that have only a handful of training surveys, make VegBench a natural testbed for future zero- and few-shot approaches, for example ones that draw on the written alliance descriptions in the *Manual of California Vegetation*.

A.4 Remote Sensing Data

We retrieved remote-sensing imagery through Google Earth Engine (GEE) using the survey coordinates as the center of each image patch. We used three imagery sources: Sentinel-2 surface reflectance from `COPERNICUS/S2_SR_HARMONIZED` (European Space Agency), Landsat (U.S. Geological Survey) Collection 2 Tier 1 Level-2 surface reflectance, and USDA National Agriculture Imagery Program (U.S. Department of Agriculture, Farm Service Agency) (NAIP) imagery from

Table A1: Number of distinct classes shared between training and each test partition at each hierarchy level, and the total number of classes in the full benchmark. The 2017+ setting restricts both training and test surveys to 2017 or later; full includes all survey years. Counts use the full training budget and exclude missing labels. The last row gives the number of surveys in each partition.

	Stratified test		Accuracy assessment test		Full
Hierarchy level	2017+	Full	2017+	Full	Benchmark
Formation class	5	6	5	6	6
Formation subclass	9	10	7	9	10
Formation	16	17	14	16	17
Division	21	22	18	21	22
Macrogroup	37	39	33	39	41
Group	65	81	66	79	86
Alliance	146	291	142	249	387
Association	172	716	122	512	1,598
CALVEG	114	207	111	182	244
CWHR	69	84	63	81	89
WHR13	27	32	28	32	35
Test surveys	548	4,900	1,499	9,421	14,321

USDA/NAIP/DOQQ. For Landsat, we used the available Collection 2 Level-2 products from Landsat 4 TM, Landsat 5 TM, Landsat 7 ETM+, Landsat 8 OLI/TIRS, and Landsat 9 OLI-2/TIRS-2, allowing the sensor contributing imagery to vary with the requested acquisition date.

For Sentinel-2, we retained the 12 surface-reflectance bands B01, B02, B03, B04, B05, B06, B07, B08, B08A, B09, B11, and B12. Candidate scenes intersecting each survey location were restricted to a temporal window around the requested date and filtered using the scene-level CLOUDY_PIXEL_PERCENTAGE metadata. We additionally masked pixels using the Sentinel-2 Scene Classification Layer (SCL), excluding classes corresponding to no data, saturated or defective pixels, cloud shadow, medium- and high-probability cloud, cirrus, and snow or ice. The remaining scenes were ordered by absolute temporal distance from the requested date, and the selected observations were combined using a per-pixel median composite.

For Landsat, we retained the blue, green, red, near-infrared, and two shortwave-infrared surface-reflectance bands from the corresponding Collection 2 Level-2 product. Candidate scenes were filtered using the scene-level CLOUD_COVER metadata. Pixel-level quality masking used the QA_PIXEL and QA_RADSAT bands to remove invalid, cloud-contaminated, or radiometrically saturated observations. As for Sentinel-2, the remaining scenes were ordered by temporal distance from the requested date and combined using a per-pixel median composite.

For NAIP, we used the GEE USDA/NAIP/DOQQ collection and selected the temporally closest acquisition intersecting each survey location rather than constructing a multi-scene composite. We retained the red, green, and blue bands and the near-infrared band when present in the source acquisition.

We constructed two temporal variants of the imagery. In the coregistered collection, the requested imagery date was the date of the corresponding vegetation survey. In the fixed collection, all survey locations used a common requested date of June 15, 2018.

A.5 geoFM Models and Baselines

A.5.1 geoFM Embedding Implementation

Here we describe how all geoFM embeddings and baselines were generated.

All geoFMs were used as frozen feature extractors with the following implementation choices:

Table A2: Input data and pretraining targets / supervision used by the evaluated representations.

Model / feature	Input source	Target / supervision source
AlphaEarth (Brown et al., 2025)	Sentinel-1, Sentinel-2, Landsat 8/9 time series	Reconstruction/prediction across inputs and ALOS PALSAR ScanSAR Level 2, GEDI L2A , ERA5-Land, GRACE, Copernicus DEM GLO-30, NLCD 2019 & 2021, Wikipedia geocoded articles, GBIF research grade observations with taxonomy text
OlmoEarth (Herzog et al., 2026)	Sentinel-1, Sentinel-2, Landsat 8/9 time series	Reconstruction/prediction across inputs and OpenStreetMap, ESA WorldCover, USDA Cropland Data Layer, SRTM DEM, canopy height, WorldCereal
TESSERA v1 (Feng et al., 2026)	Sentinel-1 SAR, Sentinel-2 time series	Self-supervised using Barlow Twins
Prithvi-EO-2.0 600M-TL (Szwarcman et al., 2025)	Harmonized Landsat–Sentinel-2 imagery, latitude/longitude, acquisition time	Masked image reconstruction
Clay v1.5 (Clay-Foundation, 2023)	NAIP or Sentinel-2 imagery, wavelength, spatial resolution (GSD), latitude/longitude, acquisition time	NAIP aerial imagery, LINZ New Zealand aerial imagery, MODIS, DINOv2 teacher supervision
DINOv3 (Siméoni et al., 2025)	RGB imagery	Self-distillation on SAT-493M (Sat) or LVD-1689M
ProM3E-Sat (Sastry et al., 2026)	Sentinel-2 embedding	Masked reconstruction across paired species iNaturalist image, iNaturalist audio, taxonomy text, lat/lon, Sentinel-2, and WorldClim-2 embeddings
TaxaBind-Sat (Sastry et al., 2025)	Sentinel-2 imagery	Contrastive alignment with paired Sentinel-2 and iNaturalist-2021 ground-level species imagery (iSatNat)
CRISP (Huynh et al., 2024)	NAIP imagery	Contrastive alignment with paired NAIP and iNaturalist ground-level plant imagery (Nature Multi-View)
Deepbiosphere (Gillespie et al., 2024)	NAIP imagery, WorldClim-2	GBIF research grade observations with genus- and family-level supervision
SatCLIP (Klemmer et al., 2025a)	Latitude/longitude	Contrastive alignment with Sentinel-2 imagery from S2-100K
MOSAIKS (Rolf et al., 2021)	RGB imagery	Unsupervised random convolutional features from randomly sampled image patches
NAIP center pixel (U.S. Department of Agriculture, Farm Service Agency)	NAIP imagery	None; direct feature
Sentinel-2 center pixel (European Space Agency)	Sentinel-2 imagery	None; direct feature
Landsat center pixel (U.S. Geological Survey)	Landsat imagery	None; direct feature
NASA WERK LiDAR (NASA Earth Exchange (NEX) et al., 2026)	30 m LiDAR-derived topography, ecosystem structure layers	None; direct features
POLARIS soil properties (Chaney et al., 2019)	30 m soil texture, organic matter, pH, hydraulic conductivity, bulk-density, water-retention properties	None; direct features
Latitude and longitude	Latitude/longitude	None; direct features
Copernicus DEM GLO-30 (2021) (European Space Agency, 2021)	30 m Copernicus GLO-30 surface elevation	None; direct feature
Latitude/longitude + Copernicus GLO-30 (European Space Agency, 2021)	Latitude/longitude, 30 m Copernicus GLO-30 surface elevation	None; direct features
ResNet50 (He et al., 2016)	RGB or multispectral Imagery	Supervised ImageNet-1K classification weight initialization
ViT-L/16 (Dosovitskiy et al., 2020)	RGB or multispectral Imagery	Supervised ImageNet-1K classification weight initialization

- **DINOv3 (Siméoni et al., 2025):** We used the DINOv3 ViT-L/16 checkpoint pretrained on LVD-1689M and used the final-layer CLS-token as the embedding. For each imagery source, only the RGB was kept, center-cropped to 224×224 pixels, and normalized using ImageNet statistics.
- **DINOv3-SAT (Siméoni et al., 2025)** We used the satellite-pretrained DINOv3-SAT ViT-L/16 SAT-493M and used the final-layer CLS-token as the embedding. For each imagery source, only the RGB was kept, center-cropped to 224×224 pixels, and normalized using the statistics released for SAT-493M.

- **CRISP (Huynh et al., 2024)** We used the released CRISP remote-sensing encoder, a four-channel ResNet-50 trained by contrastively aligning colocated NAIP and ground-level plant images, and retained its 512-dimensional aerial-image representation. To generate embeddings, NAIP tiles were center-cropped to 256×256 pixels, scaled to $[0, 1]$, and normalized with the channel statistics used during CRISP training. For historical RGB-only NAIP acquisitions, the absent near-infrared channel was filled with its training-set mean so that it was zero after normalization.
- **Deepbiosphere (Gillespie et al., 2024)** We extracted the final 2,048-dimensional layer before species prediction from Deepbiosphere, which combines a four-band NAIP-trained TResNet branch with an MLP branch trained on 19 WorldClim BioClim variables. To generate embeddings, NAIP tiles were center-cropped to a 256×256-pixel image, scaled to $[0, 1]$, and normalized per-channel with statistics from the original paper. Missing near-infrared data in historical RGB-only acquisitions were filled with the near-infrared training mean.
- **SatCLIP (Klemmer et al., 2025a)** We used the released `SatCLIP-ViT16-L40` location encoder. Embeddings were generated with lat/lon survey coordinates.
- **AlphaEarth Foundations (Brown et al., 2025)** We sampled the 64-dimensional annual AlphaEarth Foundations embedding at each survey coordinate through Google Earth Engine. Coregistered surveys used the embedding layer for the survey year, with surveys before product coverage assigned to the 2017 layer. Fixed-date surveys used the 2018 layer.
- **TESSERA v1 (Feng et al., 2026)** We sampled the hosted 128-dimensional TESSERA v1 annual representation at each survey point using GeoTESSERA. Coregistered surveys used their survey year when available and otherwise the closest published year, so pre-Sentinel surveys used 2017. All fixed-date surveys used 2018.
- **OlmoEarth v1.3 (Team OlmoEarth, 2026)** We used embeddings sampled from the Base-sized OlmoEarth v1.3 model with a bottleneck down to 128 dimensions using 2240x2240 m input imagery centered at the lat/lon survey coordinates with a 3 month export centered on June 2018 for fixed temporal inputs and a 3 month export centered on the observation date for coregistered inputs.
- **Clay v1.5 (Clay-Foundation, 2023)** We sampled precomputed Clay v1.5 embeddings hosted by Source Cooperative. For NAIP embeddings, we used the corrected `clay-v1-5-naip-2` product. Coregistered surveys used the closest available NAIP imagery year, breaking ties with the earlier year, and fixed date surveys first used NAIP 2018 imagery. For Sentinel-2 Clay embeddings, we used the 256-dimensional LGND `monthly-1280m-256d` product. Dates before product coverage used the earliest available month, and other unavailable dates used the nearest month.
- **Prithvi-EO-2.0 (Szwarcman et al., 2025)** We used the released 600M-parameter Prithvi-EO-2.0-TL encoder with Sentinel-2 median composites generating embeddings for a single timepoint. Because these inputs came from `COPERNICUS/S2_SR_HARMONIZED` rather than HLS, we selected B02, B03, B04, B8A, B11, and B12 and discarded broad near-infrared B08 (github.com/NASA-IMPACT/hls-foundation-os). The 256×256-pixel tile was center-cropped to 252×252 pixels and aggregated to an HLS-similar 84×84 grid at 30-m resolution. We then applied the released Prithvi v2 image normalization. We removed the CLS token and mean-pooled the final 6×6 spatial-token grid to obtain the final embeddings.
- **TaxaBind-Sat (Sastry et al., 2025)** We generated TaxaBind-Sat representations from Sentinel-2 RGB bands B04, B03, and B02. Reflectances were scaled to $[0, 1]$, after which images were resized, center-cropped, and normalized with ImageNet mean and standard deviation.
- **ProM3E-Sat (Sastry et al., 2026)** We used the same Sentinel-2 preprocessing and TaxaBind satellite encoder as for TaxaBind-Sat. The resulting TaxaBind-Sat vector was placed in ProM3E's satellite slot and all other modalities were marked unavailable. We retained all hidden tokens and flattened them into a 7,168-dimensional vector.

A.5.2 HAND-DESIGNED FEATURE BASELINE IMPLEMENTATIONS

For all baselines, missing and malformed input values at survey sites were filled with medians estimated from the selected training rows and paired with missingness indicators. Missing categorical values were treated as an additional category. Missing continuous values were imputed using the median value for each input.

- **NASA WERK LiDAR (NASA Earth Exchange (NEX) et al., 2026)** We sampled each of the 21 WERK canopy-height, digital-surface, and digital-terrain rasters at all survey coordinates, and any nonsense values outside the range $[-500, 9000]$ were treated as missing.
- **Copernicus GLO-30 Elevation (European Space Agency, 2021)** We used the 2021 Copernicus DEM GLO-30 mosaic at each survey coordinate to infer elevation.
- **POLARIS Soil Properties (Chaney et al., 2019)** We used the POLARIS mean estimates for clay, sand, silt, pH, organic matter, bulk density, saturated hydraulic conductivity, saturated water content, α, and n at depths of 0–5, 5–15, 15–30, and 30–60 cm.
- **Latitude and Longitude** The spatial baseline consisted of the authoritative survey longitude and latitude in WGS84 coordinate reference system.
- **Remote Sensing Imagery Center Pixels** We constructed sensor-specific spectral baselines by using the center pixel of every band for NAIP, Sentinel-2, and Landsat. For an even-sized raster, the selected pixel was $(\lfloor h/2 \rfloor, \lfloor w/2 \rfloor)$. Values were used in their raw reflectance units without scaling.

A.5.3 MOSAIKS (ROLF ET AL., 2021)

We implemented the original empirical MOSAIKS random-feature construction on native-resolution NAIP. The filter bank was learned from 100,000 randomly sampled $3 \times 3 \times 3$ RGB patches and reduced to 4,096 filters by uniform sampling without replacement. Patch preprocessing used the released normalization and ZCA transform with bias 0.001, including the two applications of the symmetric inverse-square-root ZCA matrix in the released v1.0 code. A fixed random seed of zero was used. To avoid benchmark leakage, candidate filter-bank chips were sampled outside a 3,840-m buffer around every survey coordinate. One native-resolution 256×256-pixel chip was sampled from each eligible NAIP tile, after eroding tile footprints sufficiently to keep chips within the raster boundary. Survey features were then computed from native 512×512-pixel NAIP RGB crops. We globally averaged positive and negative convolutional responses separately, producing 8,192 features per survey.

A.5.4 RESNET AND VIT BASELINE TRAINING

ResNet-50 and ViT-L/16 are initialized from ImageNet-1K weights and fine-tuned end to end on RGB image tiles center-cropped to 224×224 pixels, with a separate model for each input source and hierarchy level. Both use AdamW for up to 30 epochs with early stopping (patience 5), keeping the checkpoint with the highest validation macro-F1. ResNet-50 uses a batch size of 32, learning rate 10^{-4} with cosine decay, and weight decay 10^{-4}. ViT-L/16 uses a batch size of 16, a constant learning rate of 10^{-5}, weight decay 10^{-2}, and label smoothing of 0.1. These settings maximized mean validation macro-F1 in a preliminary hyperparameter sweep. Results average three random seeds.

A.5.5 LINEAR PROBE HYPERPARAMETER SELECTION

For every frozen representation and hand-designed feature set, we standardize features using training-set statistics and fit an L2-regularized multinomial logistic regression (L-BFGS, up to 2,000 iterations), with a separate probe for each hierarchy level. The inverse regularization strength is selected from $C \in \{0.01, 0.1, 1, 10, 100\}$ by validation macro-F1.

A.6 Extended Results

A.6.1 How does balanced accuracy change along the hierarchy?

Similar to standard accuracy results in Fig. 3, balanced accuracy in Fig. A1 is higher for time series geoFMs and NAIP image-based embeddings, but decreases as the vegetation hierarchy gets finer. Overall, performance is skewed lower compared to overall accuracy as the less common and thus harder to identify alliances are have a higher weighting in balanced accuracy calculation.

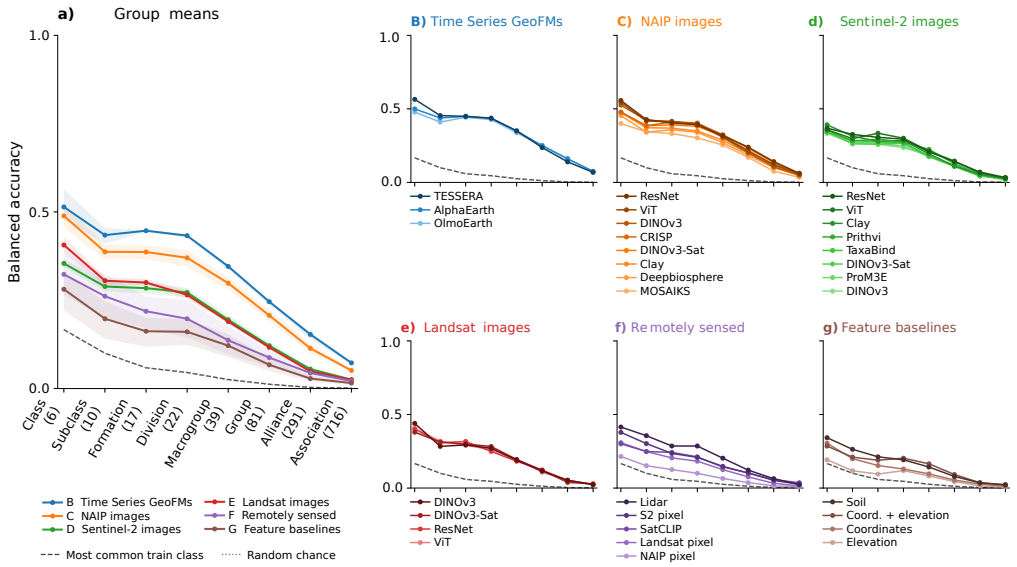

Figure A1: **GeoFMs perform similarly for balanced accuracy compared to standard accuracy along the vegetation hierarchy.**

A.6.2 How does sensor resolution affect model accuracy?

In Fig. A2, geoFM embeddings and remotely sensed baselines tested with both NAIP, Sentinel-2, and Landsat imagery perform on average overwhelmingly better at predicting vegetation alliance from NAIP imagery than either Sentinel-2 or Landsat imagery. NAIP-based embeddings perform nearly twice as well as Sentinel-2 and more than double Landsat imagery for most geoFMs. This trend holds for models trained end-to-end on imagery (ResNet50, ViT-Large), trained on both modalities (Clay), and for which imagery is out of the training set distribution (DinoV3, DinoV3-Sat). Only for the center pixel baseline is NAIP imagery worse, likely due to the few number of bands and high reflectance variance for individual pixels.

A.6.3 How does accuracy change when classes are spectrally mixed?

We call a survey *pure* when the expert map assigns the same alliance to the entire 135 m NAIP tile around it, and *mixed* otherwise, and compare accuracy on the two groups of accuracy assessment surveys (Fig. A3). Every model is more accurate on pure surveys (e.g., AlphaEarth 48% vs. 27% alliance accuracy), but most of this difference reflects which vegetation types tend to occur in uniform patches: the location-only baseline, which sees no imagery, shows a similar gain (21% vs. 9%). Comparing pure and mixed surveys within the same alliance (55 alliances with both), the advantage shrinks to 3–8 percentage points for most imagery-based models and disappears for the location-only baseline. Model rankings are nearly identical on pure and mixed surveys (Spearman $\rho \geq 0.96$), so spectral mixing within the image tile does not explain the differences between models reported in the main text. In relative terms, Sentinel-2 and Landsat models benefit most from purity (roughly twice the accuracy on pure surveys, versus 1.7 times for time series geoFMs), consistent with their coarser pixels mixing neighboring vegetation.

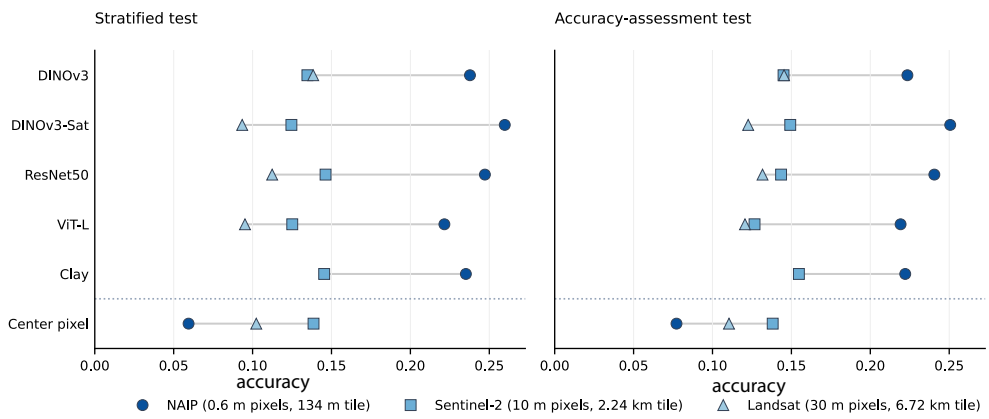

Figure A2: **High resolution NAIP imagery-based embeddings overwhelmingly outperform Sentinel-2 and Landsat-based approaches for predicting vegetation alliance**. Accuracy for embeddings from five single-image geoFMs compared across three different sensor modalities. Both Sentinel-2 and NAIP data is in training distribution for Clay, plus Landsat for ViT-L and ResNet50. Sentinel-2 and Landsat are out of distribution for DinoV3 and DinoV3-Sat.

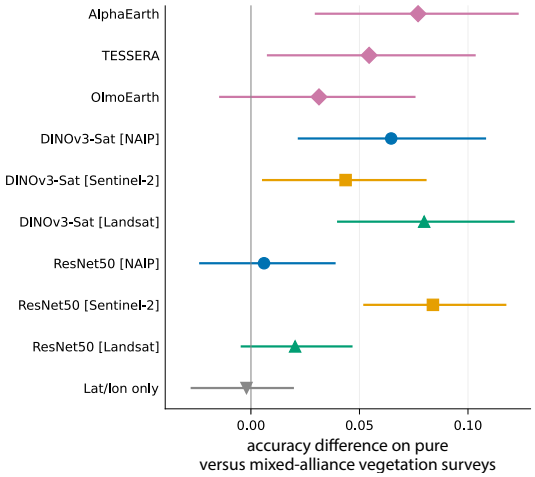

Figure A3: **GeoFMs generally predict vegetation alliance better using surveys where vegetation alliances are pure and not mixed.** Since vegetation surveys are often fixed-point rapid assessments or relevés, the entire remotely-sensed image or local region used to generate embeddings may not belong to the same vegetation alliance. For the 55 alliances with both accuracy assessment surveys and a corresponding expert vegetation map, accuracy on "pure" surveys where 100% of the 135x135 m region surrounding a survey is that same vegetation alliance, versus not ("mixed-alliance" surveys) tends to be significantly higher (bars 90% CI over considered alliances).

A.6.4 How do results differ across co-registration design choices?

Across the majority of models, in Fig. A4 using all the historical surveys aligned at a fixed date produced the best results (filled circle), consistent with trends in Fig. 4b that more training surveys produces better vegetation classification models. Interestingly, co-registering embeddings by year of vegetation survey acquisition for surveys from 2017 and later (orange diamond) overwhelmingly is worse than using a fixed date of 2018 for extracting embeddings (hollow circle). Given that the VegCAMP surveys underpinning the VegBench benchmark are collected per-project, surveys collected in the same year tend to spatially cluster, introducing a strong spatiotemporal autocorrelation in the data. In contrast, vegetation type changes slowly over time, on the order of decades. Therefore, autocorrelation-driven overfitting is more likely to influence classifier behavior than are mismatches in surveyed versus sensed vegetation driven by temporal mis-registration. We include

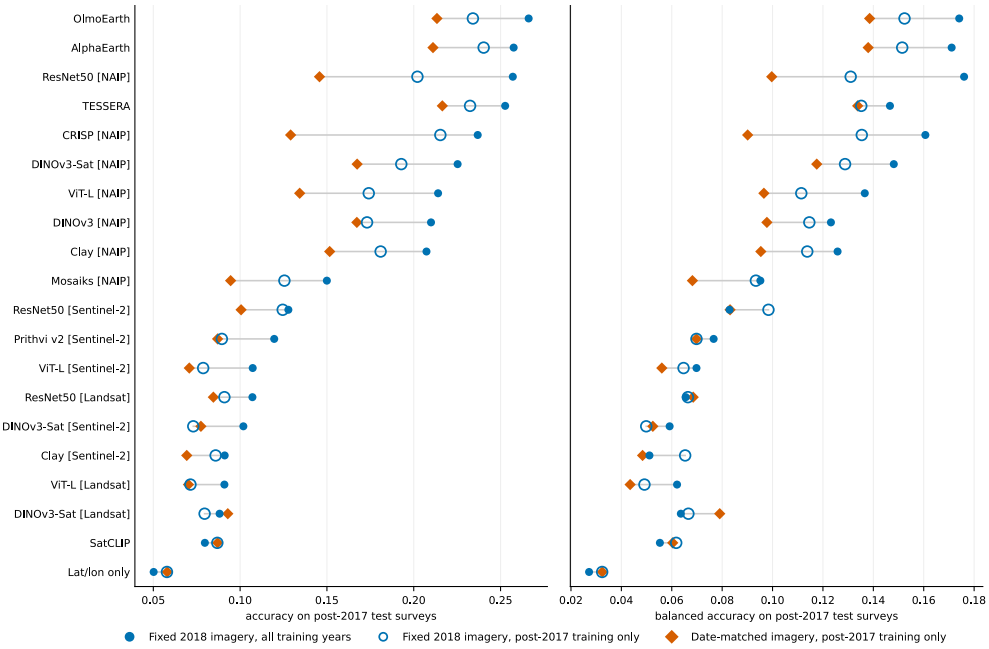

Figure A4: **GeoFMs better predict vegetation alliance with more training data, and worse when embeddings and surveys are temporally co-registered.** Fitting probes with all historical surveys overwhelmingly leads to better performance (filled circle) except for location-based baselines. In contrast, temporally co-registered surveys and embeddings by year for post-2017 vegetation surveys (orange diamond) leads to generally worse performance than using a fixed year for all embeddings (2018, hollow circle).

SatCLIP and Lat/Lon baselines in Figure A4 to show variation between the larger and smaller training sets, since they do not consider temporal information in the modeling process. We see that these models usually perform worse than the other models across metrics and test sets and have relatively small differences between different training set sizes.

A.6.5 Does stratification versus accuracy assessments make a difference?

Spatially buffering vegetation surveys to ensure minimal overlap between the test and train satellite tiles in Fig. A5 does not appear to lead to significant differences in geoFM performance compared to using accuracy assessment surveys. The best performing embeddings are slightly more accurate on spatially stratified compared to accuracy assessment surveys, but differences are small (average accuracy difference: 1.3%, balanced accuracy: 1.4%) and do not materially affect overall model ordering (Spearman's ρ for accuracy = .97, balanced accuracy = .91).

A.6.6 Does probe choice affect classifier performance?

We compared accuracy for a subset of geoFMs across four downstream classification strategies in Fig. A6: fitting a linear probe as reported in the rest of the manuscript, a multilayer perceptron (MLP), a random forest, and imputing vegetation alliance using nearest neighbors. Relative accuracy between model groupings changed little across probe types. Random forest had similar performance to a linear probe, while MLP was slighty lower accuracy, and nearest neighbor imputation even lower. Sentinel-2-based embeddings performed better for nearest neighbors imputation compared to other probes, likely due to the larger overall embedding dimensionality of the Sentinel-2-based models, but not enough to change general geoFM performance rankings. For MLP specifically, Sentinel-2 models performed especially poorly, again likely due to the larger average embedding dimension size of the Sentinel-2 models.

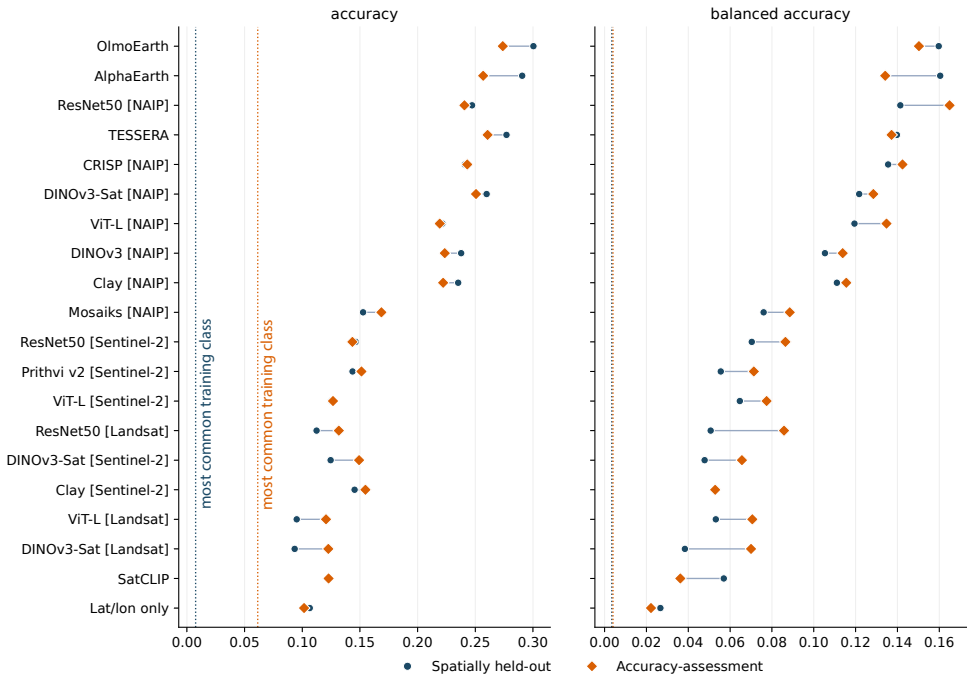

Figure A5: **GeoFM performance is similar across test splitting strategies.** The best-performing geoFMs like OlmoEarth perform slightly better when using spatially stratified test surveys, while lower-accuracy models, including fully supervised image baselines, perform slightly better on accuracy assessment test surveys. However, aggregate trends across sensor groupings remains similar and accuracy differs $< 3\%$ on average.

A.6.7 How do draft vs. published expert maps compare to geoFM performance?

As part of VegCAMP's standardized vegetation mapping workflow, after expert maps are created and accuracy assessment data collected, the maps' accuracy are tested against the accuracy assessment data. In the case that an expert map exhibits certain systematic errors, the mapping team will occasionally manually edit the map using the collected accuracy assessment data to correctly relabel alliances. While the preliminary map accuracies pre-edits are not systematically published by VegCAMP, for the Southern Sierra Nevada Foothills vegetation mapping project, the final map report (Reyes et al., 2022) lists the pre-edit map hierarchical accuracy, which we use to compare to the final released expert vegetation map accuracy in Fig. A7. The manual map editing resulted in an expert map accuracy increase of only 5%, a marginal difference compared to the $> 45\%$ gap between the unedited expert map and the best average geoFM grouping. Although expert maps may incorporate information from accuracy assessment surveys (i.e. our test locations), not all are modified to align with ground surveys; that is, there is still a residual error in the publicly released data, which we report here and in the main text. This discrepancy may be a result of operational or economic constraints, and we also stress that the unaudited polygons in the survey still carry some classification uncertainty which we assume is of a similar magnitude to the result of the accuracy assessment.

A.6.8 How does geoFM performance scale with number of training surveys?

Comparing accuracy for alliances with good test set coverage (≥ 5 test surveys) and that are possible to predict from remotely sensed imagery (expert map alliance accuracy $\geq 69\%$), in Fig. A8 we see that even for alliances that are easily classified by expert maps, geoFMs still struggle ($\sim 20\%$ on average for time series geoFMs, $\sim 17.5\%$ on average for NAIP image models). GeoFMs do show a strong positive relationship between number of training surveys and accuracy, suggesting more ground survey data collection effort would result in better geoFM vegetation classification performance. However, given that in aggregate collecting these vegetation surveys costs roughly

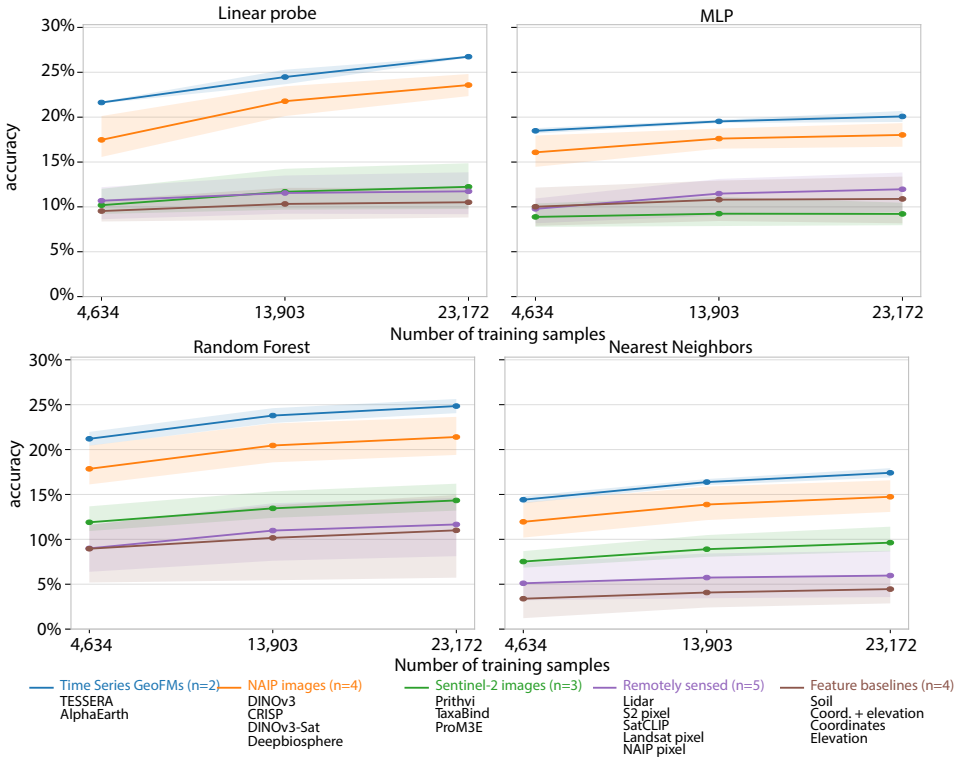

Figure A6: **Downstream classifier strategy changes absolute but not relative performance between geoFMs.** Random forest classifier shows similar accuracy to a linear probe, while MLP and nearest neighbors imputation are overall lower accuracy. Relative ordering between geoFM groups remains stable across both probes and number of training samples.

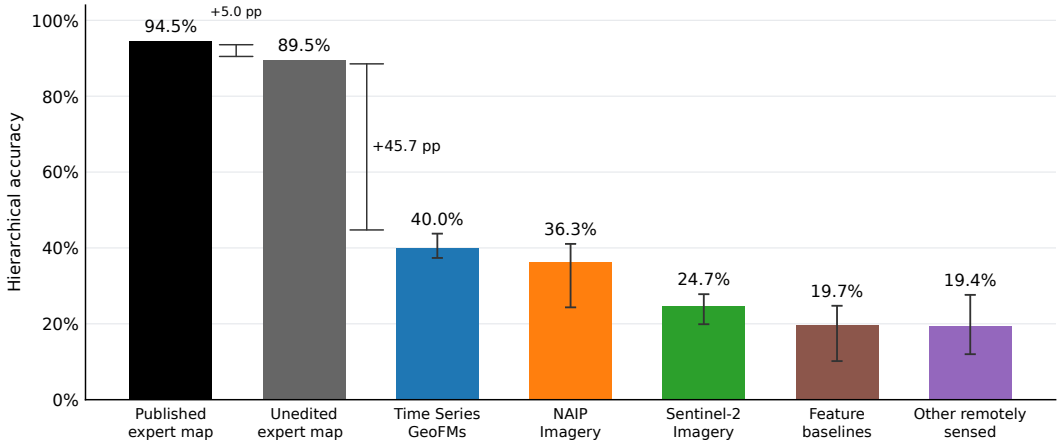

Figure A7: **Unedited expert vegetation maps are much closer in accuracy to published expert maps than the next-best geoFM embeddings.** For the Southern Sierra Nevada Foothills vegetation mapping project (Reyes et al., 2022), we compare unedited versus published expert vegetation map accuracy to our suite of evaluated models.

$1,000 per-survey (Ratchford et al., 2026), collecting ∼1,000 surveys for 300+ vegetation alliances would cost upwards of USD $300 million.

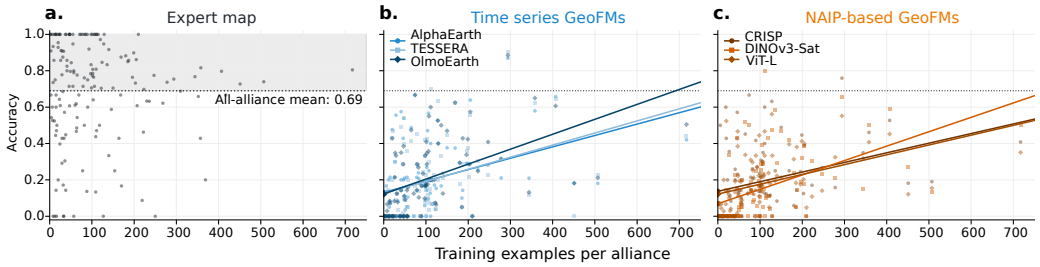

Figure A8: **GeoFM performance can scale to expert map performance, but only with far many more examples per-alliance. a.** Per-alliance accuracy for alliances with ≥ 5 accuracy assessment surveys falling within alliance-resolvable expert-map polygons. To ensure alliances are resolvable from remotely-sensed imagery, for **b.** time series and **c.** NAIP-based geoFMs we only consider geoFM performances for alliances above the expert map average alliance accuracy (accuracy = 0.69, shaded region of expert map). Extrapolating linearly, on average per-alliance time series geoFMs would need 865 and NAIP image models 1,062 examples to reach the average expert map alliance classification performance of 0.69.

A.6.9 Does crosswalking coarser vegetation categories affect performance?

Given the hierarchical nature of vegetation classification, for coarser levels of the hierarchy like group, macrogroup, and formation, classifiers can either be fit to directly predict those labels, or crosswalked from finer-scale vegetation type classifiers, e.g. alliance probes. Comparing crosswalked versus directly-trained vegetation classification accuracy across three levels of the hierarchy, in Fig. A9 we see that for time series geoFMs crosswalking accuracy is no different than direct training accuracy. In contrast, for the NAIP, Sentinel-2, Landsat, and baseline models direct training accuracy is significantly higher. The performance boost is consistent across sensor modalities, but is still smaller than the relative accuracy gap between sensor modalities. This suggests that geoFM performance at higher levels of the hierarchy can be confidently evaluated using a single trained probe for alliance-level prediction.

A.6.10 Can geoFMs accurately predict land use at VegBench surveys?

To enable direct comparison of VegBench surveys to standard geospatial benchmarks, we also evaluate test accuracy for remote sensing-resolvable vegetation categories (CalVeg), coarse California habitat categories (CWHR), and standard land use categories (WHR13) in Fig. A10. Broadly speaking, as the number of categories decreases, geoFM accuracy increases, recapitulating trends seen across the vegetation hierarchy (Fig. 3). Relative performance differences between geoFMs also largely hold across class prediction targets. Accuracy for prediction targets like WHR13 that are closest to standard land use / land cover geospatial benchmark classification tasks like GLaNCE are highly similar (e.g. AlphaEarth (Brown et al., 2025) GLaNCE accuracy = 0.68 vs. VegBench WHR13 accuracy of 0.67). This similar accuracy, plus the linear relationship between alliance and coarser classes, implies improved VegBench performance will also lead to better general-task geoFM behavior for other standard geospatial benchmarks.

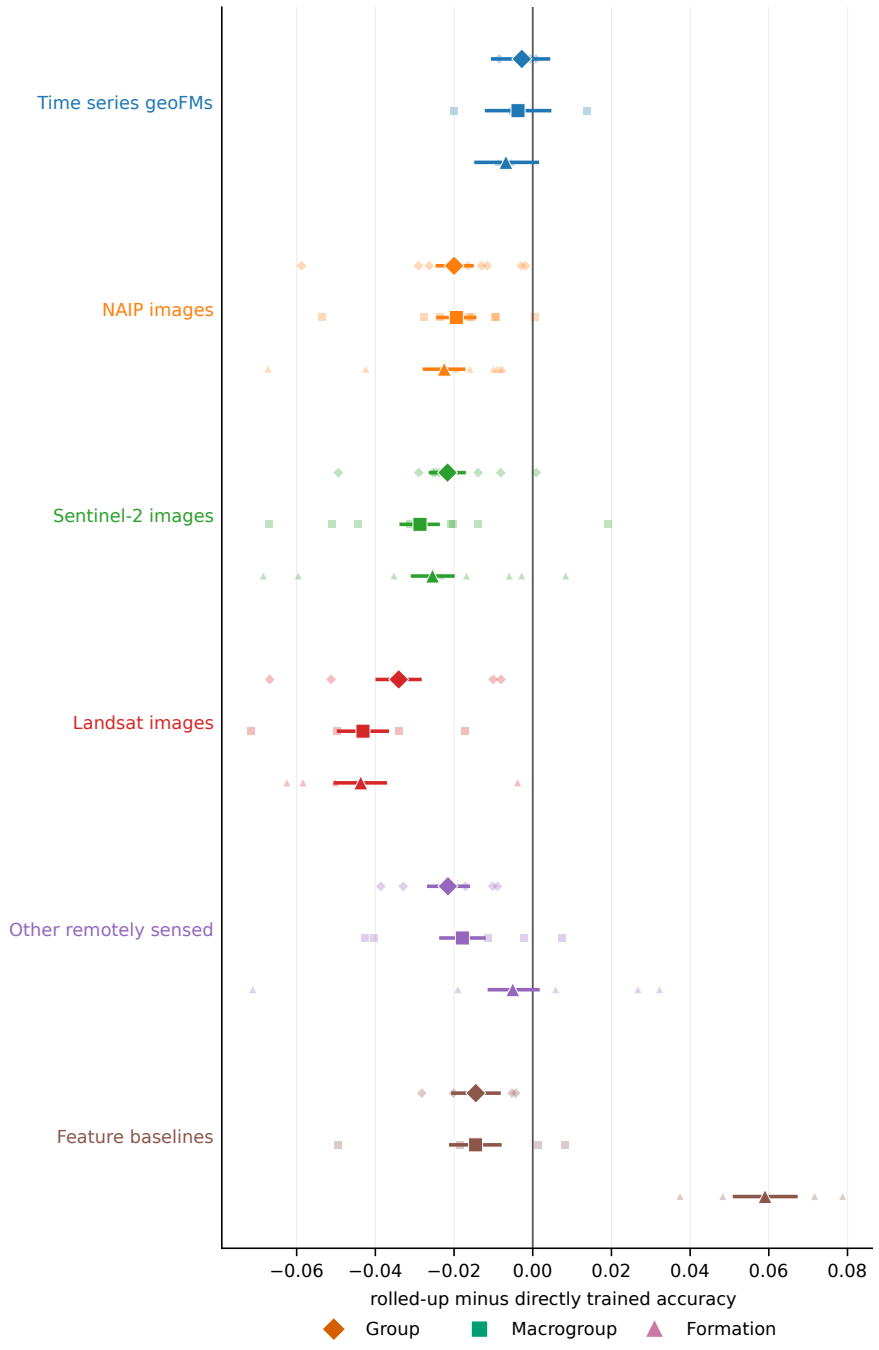

Figure A9: **Accuracy differences for higher vegetation classes using alliance-trained and crosswalked versus direct coarse class training is minimal.** Training probes to directly predict group, macrogroup, and formation versus crosswalking an alliance-trained probe to coarser vegetation categories confers no significant advantage for time series geoFMs. Meanwhile, direct training helps for NAIP, Sentinel-2, Landsat, and baseline models, but the difference is still smaller than the accuracy gap between time series geoFMs and these sensor modalities (Fig. 3).

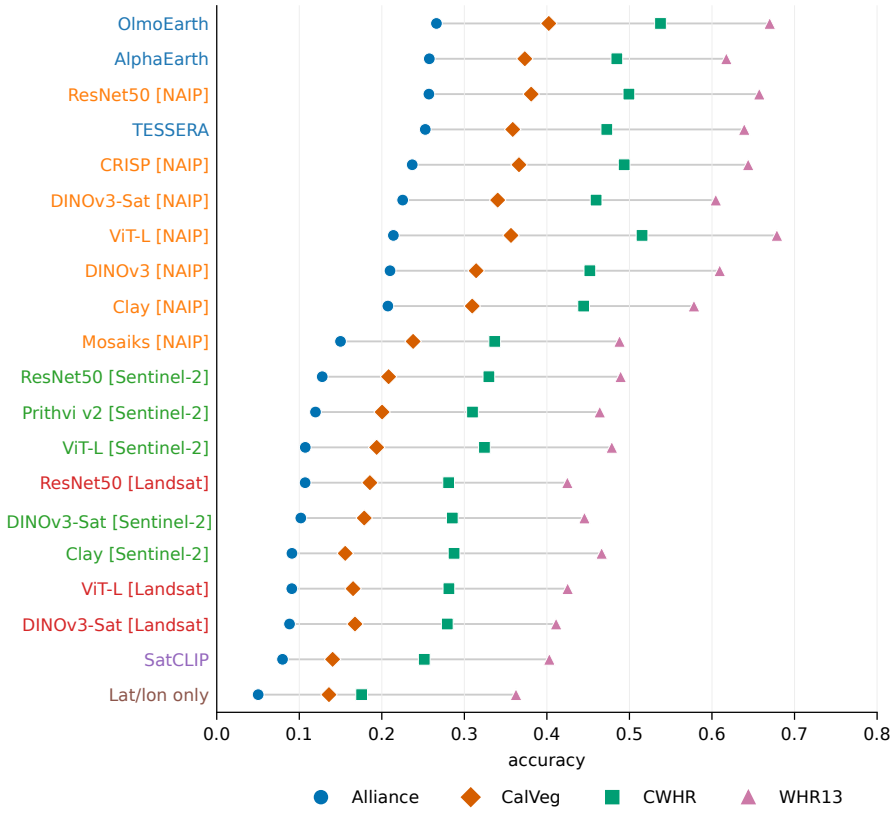

Figure A10: **GeoFM embedding accuracy for predicting land use & land cover classes at VEGBENCH test survey sites is on par with other geospatial benchmarks.** For remote sensing-resolvable vegetation categories (CalVeg), coarse California habitat categories (CWHR), and standard land use categories (WHR13), geoFM accuracy is significantly higher across the board predicting these coarser classes than alliance class prediction. Performance patterns also largely hold across models, with better-performing geoFMs for vegetation alliance prediction also performing better for land use categorization.