A. Research Benchmarks

We describe the research benchmarks introduced in Section 3.2 in more detail below. We also share our observations on limitations of certain benchmarks.

GEO-Bench modifies benchmarks to form a unified and consistent collection of datasets:

m-bigearthnet is modified from BigEarthNet [47], which involves multi-label land cover classification of 120×120 Sentinel-2 image crops. It consists of 19 classes, such as arable land, inland wetlands, and urban fabric. The original dataset contains 549,488 examples, but the modified subset in GEO-Bench contains only 22,000, with 20,000 for training, 1,000 for validation, and 1,000 for testing.

m-so2sat is modified from So2Sat LCZ42 [65], which involves image-level classification of local climate zones from co-registered Sentinel-1 and Sentinel-2 crops. It consists of 17 classes, such as high-rise, industrial, and water bodies. The original dataset contains 400,673 examples, but the modified subset in GEO-Bench contains only 21,964, with 19,992 for training, 986 for validation, and 986 for testing.

m-brick-kiln is modified from the Brick Kiln Classification Dataset in Bangladesh [33]. The original dataset involves image-level classification of whether or not high-resolution 224×224 satellite image crops from DigitalGlobe contain at least one kiln, and contains 6,329 positive examples and 67,284 negative examples. The modified dataset in GEO-Bench performs the same task on corresponding 64×64 Sentinel-2 crops, and contains only 17,061 examples, with 15,063 for training, 999 for validation, and 999 for testing. While finding kilns in Sentinel-2 images is a challenging task, we find that the nature of the negatives in the GEO-Bench version of the dataset make the classification task too easy; for example, many negatives seem to have only dark pixels, making it easy to distinguish them.

m-forestnet is modified from ForestNet [26], which involves image-level classification of deforestation drivers from a composite 332×332 Landsat 8 satellite image captured within five years after each forest loss event. There are four driver categories: plantation, smallholder agriculture, grassland/shrubland, and other. The original dataset contains 2,756 examples. The modified subset in GEO-Bench contains 6,464 examples for training, 989 examples for validation, and 993 examples for testing; we could not determine where the additional examples came from.

m-eurosat is modified from EuroSat [24], which involves image-level land use and land cover classification from 64×64 Sentinel-2 image crops. It consists of 10 classes, such as annual crop, river, and highway. The original dataset contains 27,000 examples, but the modified subset in GEO-Bench contains only 4,000, with 2,000 for training, 1,000 for validation, and 1,000 for testing.

m-cashewplant is modified from the Smallholder Cashew Plantations in Benin Dataset [29], which involves segmentation of 256×256 Sentinel-2 image crops. It consists of six classes relating to cashew plantations: well-managed plantation, poorly-managed plantation, non-plantation, uncertain, residential, and background. The modified dataset contains 1,800 examples, with 1,350 for training, 400 for validation, and 50 for testing. Multiple models were sensitive to input patch size on this dataset, so for models that had a variable patch size, we swept input patch size and report the best result. Ultimately this is likely an effect of the labels being large polygons instead of per-pixel labels.

m-SA-crop-type is modified from the South Africa Crop Type Competition dataset [60], which involves crop type segmentation of 256×256 Sentinel-2 and Sentinel-1 image crops. It consists of 10 classes, such as fallow, wine grapes, and wheat. The modified dataset in GEO-Bench only uses the Sentinel-2 images, and contains 5,000 examples, with 3,000 for training, 1,000 for validation, and 1,000 for testing.

All of the GEO-Bench datasets share a significant limitation: although the tasks involve labels that do not change rapidly over time, the input consists of a single satellite image or image pair. We find that remote sensing models generally perform much better with multiple input images, and argue that single-image inputs should only be used for tasks like vessel detection where the labels are only valid for one timestep.

We compare on five additional datasets outside of GEO-Bench:

BreizhCrops [44] involves crop type classification from single-pixel Sentinel-2 time series. It consists of nine classes, such as wheat, corn, and permanent meadows. It contains 610K examples.

CropHarvest [51] involves binary cropland classification from single-pixel time series. The provided time series include Sentinel-2 and Sentinel-1 satellite image observations, as well as elevation from SRTM and weather data from ERA-5. It contains 95,186 examples.

PASTIS [17] involves crop type segmentation from Sentinel-1 and Sentinel-2 image time series, with 128×128 image crops. It consists of 19 classes, such as grapevine, spring barley, and soybeans. It contains 2,433 examples.

MADOS [31] involves marine debris segmentation in 80×80 Sentinel-2 image crops. It consists of 15 classes, such as oil spills, dense sargassum, and foam. It contains 2,803 examples. A key limitation with MADOS is that it provides custom-processed images, making it difficult to apply foundation models with their intended normalization statistics. Additionally, the dataset includes a lot of rare classes that greatly affect mIoU in the test set, making metrics highly variable across runs of the same model with dif-

ferent seeds.

Sen1Floods11 involves binary water segmentation in 512×512 Sentinel-2 image crops that focus on flooded areas. It contains 4,831 examples. All of the remote sensing models we tested get between 78-80% accuracy, and we find that the accuracy is not well correlated with other benchmarks. However, Sen1Floods11 is one of the few Sentinel-1 benchmarks.

B. Partner Tasks

We describe the partner tasks introduced in Section 3.2 in more detail below.

AWF - *African Wildlife Foundation (AWF)* Land cover classification in southern Kenya. The dataset contains 1,459 examples with 9 classes, which range from lava forest and agriculture to urban development. The AWF team used Planet imagery as the main reference to annotate these examples.

Live Fuel Moisture Content - *NASA JPL* Regression dataset of 41,214 examples from Globe-LFMC-2.0 [63] labeled with the LFMC value. We partner with NASA JPL to deploy a model trained on this data. LFMC predictions are used to understand wildfire risk.

Mangrove - *Global Mangrove Watch* Classification dataset of 100,000 coastal areas into 3 classes: mangrove forest, water, or other. Mangrove maps across different years are used to understand mangrove growth and loss.

Nandi - *CGIAR* Crop-type classification in Nandi County, Kenya. The dataset contains 6,924 examples with 6 categories (coffee, maize, sugarcane, etc.). The ground-truth labels were collected through field surveys.

Ecosystem type mapping is similar, but only uses six timesteps of input images:

GEA North Africa - *Global Ecosystem Atlas* Ecosystem type classification of 2,361 examples in a region of North Africa, and labels correspond to the 110 categories in level 3 of the IUCN Global Ecosystem Typology [20].

The other tasks are more unique:

Forest Loss Driver - *Amazon Conservation* Classification dataset for the cause of forest loss in the Amazon rainforest into 10 classes (mining, logging, agriculture, etc.). The input consists of 4 Sentinel-2 images captured before the forest loss and 4 images captured after the forest loss. Driver predictions are used to prioritize enforcement and litigation efforts to deter further human-caused forest loss.

Marine Infrastructure - *Skylight* Global marine infrastructure detection dataset containing 7,197 examples labeled as offshore platform or wind turbine. The input consists of a time series of 4 Sentinel-2 or Sentinel-2 + Sentinel-1 images.

Vessel Detection, Type, Length - *Skylight* Three object detection tasks to detect vessels in Landsat (8,000 examples), Sentinel-1 (1,776 examples), and Sentinel-2 (45,545 examples) images, one classification task to predict the vessel type in Sentinel-2 images centered at detected vessels (584,432 examples), and one regression task to estimate the vessel length in Sentinel-2 images (584,432 examples). For all of these tasks, the input is a single image.

Solar Farm Detection: Binary segmentation dataset containing 3,561 examples densely labeled with solar farm polygons. The input consists of 4 timesteps, either Sentinel-2 or Sentinel-2 + Sentinel-1. Solar farm maps are used to understand the global rate of renewable energy deployment over time.

C. Linear Probing Results across Random Seeds

For each model and benchmark task where we report linear probing results, we train a linear probe using 10 random seeds. We use the learning rate and normalization settings that produced the highest validation score with the seed run for the main experiments in the paper to limit the total number of experiments needed to run. The probe with the max validation score over 50 epochs is used to evaluate the results on the test set. The values in Table 6 are the mean and 95% confidence interval (CI) across seeds. Confidence intervals are computed using the Student's t-distribution over the per-seed test metrics for each (model, task) pair. Across seeds, most tasks show tight CIs, indicating stable linear probing. CropHarvest tasks show greater variability.

D. Fine-tuning Results across Random Seeds

For each model and research benchmark task, we fine-tune using three random seeds. For a given seed, we sweep learning rates over $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, and select the checkpoint with the highest validation score and report the test performance from that checkpoint as the seed's results. The values in Table 7 are the mean and 95% confidence interval (CI) across seeds. Confidence intervals are computed using the Student's t-distribution over the per-seed test metrics for each (model, task) pair. Across seeds, most tasks show tight CIs, indicating stable fine-tuning. However, a few tasks, particularly m-so2sat, m-forestnet, and MADOS, show greater variability.

E. Additional Ablations

In addition to the ablations in Section 3.6, we conduct a second set of ablations in Table 8. Our second set of ablations evaluates the contributions of components of our final model and training recipe by removing them individually, with the exception of the top row which is a MAE baseline. These models are trained for 300,000 steps. In the data ablation section we see the Sentinel-2 only model perform relatively poorly, however the "No Maps" run (only observational data) maintains relatively high performance.

		BreizhCrops	m-cashewplant	m-SA-crop-type	CropHarvest-PRC	CropHarvest-PRC	CropHarvest-PRC	CropHarvest-Togo	CropHarvest-Togo	CropHarvest-Togo	PASTIS	PASTIS	PASTIS	MADOS	Sen1Floods11
	Modalities	S2	S2	S2	S2	S1	S1+S2	S1	S2	S1+S2	S1	S2	S1+S2	S2	S1
	Time series	✔	✘	✘	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Model	Metric	Acc.	mIOU	mIOU	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	mIOU.	mIOU	mIOU	mIOU	mIOU
Anysat	ViT Base	57.8 ± 3.3	24.1 ± 0.3	17.0 ± 6.0	75.3 ± 1.6	57.7 ± 0.7	72.0 ± 0.9	73.1 ± 1.7	76.3 ± 1.6	73.7 ± 2.3	23.2 ± 0.6	38.9 ± 1.7	39.1 ± 1.8	41.0 ± 0.2	77.9 ± 0.1
Clay	ViT Large	54.1 ± 0.1	31.9 ± 0.3	23.1 ± 0.1	65.6 ± 0.4	56.1 ± 0.2	62.7 ± 0.6	77.1 ± 1.5	67.7 ± 2.3	70.4 ± 2.3	20.0 ± 0.1	22.7 ± 0.1	24.3 ± 0.1	47.8 ± 0.1	78.8 ± 0.1
CopernicusFM	ViT Base	65.7 ± 0.3	32.1 ± 0.1	28.3 ± 0.1	72.6 ± 0.7	55.7 ± 0.7	72.1 ± 1.7	77.8 ± 0.6	76.2 ± 1.5	74.4 ± 1.2	15.8 ± 0.0	32.2 ± 0.1	29.8 ± 0.1	63.5 ± 0.2	77.6 ± 0.2
CROMA	ViT Base	68.3 ± 1.6	24.5 ± 0.3	24.3 ± 2.5	75.2 ± 0.7	56.0 ± 0.9	73.7 ± 1.0	77.5 ± 1.3	78.3 ± 1.4	78.9 ± 1.6	24.8 ± 0.6	44.7 ± 0.1	38.3 ± 0.3	60.8 ± 0.7	78.2 ± 0.2
CROMA	ViT Large	66.7 ± 1.0	26.4 ± 0.2	26.2 ± 1.4	71.6 ± 1.1	56.2 ± 0.5	73.3 ± 1.3	77.5 ± 1.1	79.6 ± 1.6	77.8 ± 1.4	25.0 ± 0.3	42.6 ± 0.0	37.6 ± 0.2	66.0 ± 0.1	78.8 ± 0.1
Panopticon	ViT Base	57.7 ± 0.1	32.8 ± 0.0	27.4 ± 0.1	74.8 ± 0.9	57.1 ± 0.9	75.9 ± 0.6	73.3 ± 1.6	75.4 ± 1.3	75.0 ± 1.2	23.2 ± 0.0	30.3 ± 0.0	29.7 ± 0.0	66.1 ± 0.1	78.1 ± 0.1
Prithvi v2	ViT Large	57.6 ± 0.6	45.2 ± 0.1	27.4 ± 0.1	71.5 ± 0.5	–	–	–	74.8 ± 2.4	–	–	34.6 ± 0.1	–	56.6 ± 0.2	–
Prithvi v2	ViT Huge	58.9 ± 0.7	42.9 ± 0.1	29.7 ± 0.1	72.3 ± 0.5	–	–	–	73.2 ± 2.3	–	–	35.3 ± 0.1	–	57.6 ± 0.1	–
Satlas	Swin Base	55.6 ± 0.3	27.4 ± 0.0	25.7 ± 0.1	70.8 ± 0.3	64.2 ± 0.4	–	77.7 ± 0.9	79.3 ± 0.5	–	10.8 ± 0.0	14.3 ± 0.0	–	32.1 ± 0.1	71.6 ± 0.1
TerraMind	ViT Base	65.0 ± 0.3	41.5 ± 0.1	30.6 ± 0.1	76.0 ± 1.4	56.6 ± 0.6	76.5 ± 0.6	74.5 ± 2.3	71.7 ± 2.7	77.5 ± 1.3	22.6 ± 0.1	40.9 ± 0.1	39.6 ± 0.1	65.9 ± 0.6	78.8 ± 0.1
TerraMind	ViT Large	65.5 ± 0.2	45.6 ± 0.1	31.4 ± 0.1	76.4 ± 0.8	56.8 ± 0.7	77.7 ± 0.9	76.1 ± 1.9	77.9 ± 1.1	76.3 ± 1.6	22.3 ± 0.0	41.4 ± 0.1	39.9 ± 0.0	67.3 ± 0.1	78.7 ± 0.1
DINOv3 Sat	ViT 7B	31.3 ± 0.0	54.0 ± 0.1	31.6 ± 0.1	71.7 ± 1.1	–	–	–	67.7 ± 1.4	–	–	26.1 ± 0.1	–	59.9 ± 0.3	–
DINOv3 Sat	ViT Large	31.3 ± 0.0	32.4 ± 0.1	28.6 ± 0.1	70.4 ± 0.7	–	–	–	68.2 ± 3.3	–	–	22.0 ± 0.1	–	57.1 ± 0.3	–
DINOv3	ViT 7B	31.3 ± 0.0	34.0 ± 0.0	28.0 ± 0.1	70.1 ± 1.2	–	–	–	71.6 ± 1.7	–	–	21.3 ± 0.1	–	52.2 ± 0.1	–
DINOv3	ViT Base	31.3 ± 0.0	23.6 ± 0.1	26.8 ± 0.1	64.5 ± 1.6	–	–	–	65.2 ± 2.0	–	–	18.2 ± 0.1	–	53.5 ± 0.1	–
DINOv3	ViT Huge	31.3 ± 0.0	25.0 ± 0.1	26.7 ± 0.1	68.6 ± 1.1	–	–	–	70.5 ± 1.4	–	–	17.4 ± 0.1	–	48.4 ± 0.2	–
DINOv3	ViT Large	31.3 ± 0.0	24.6 ± 0.1	26.4 ± 0.1	64.5 ± 1.0	–	–	–	68.7 ± 1.4	–	–	17.4 ± 0.1	–	52.5 ± 0.4	–
Galileo	ViT Nano	58.3 ± 0.9	22.0 ± 0.2	18.6 ± 0.2	73.0 ± 1.3	60.1 ± 1.1	75.4 ± 0.8	75.1 ± 0.7	70.7 ± 1.2	78.0 ± 0.6	18.6 ± 0.2	18.8 ± 0.2	18.6 ± 0.2	53.1 ± 0.3	78.6 ± 0.0
Galileo	ViT Tiny	62.2 ± 0.8	23.6 ± 0.4	21.5 ± 0.2	78.4 ± 1.3	58.1 ± 0.7	79.0 ± 0.6	77.1 ± 1.0	76.5 ± 1.0	77.1 ± 0.7	21.8 ± 0.2	27.1 ± 0.1	24.6 ± 0.0	60.9 ± 0.1	78.6 ± 0.1
Galileo	ViT Base	66.8 ± 0.2	24.6 ± 0.4	25.3 ± 0.0	77.7 ± 0.7	63.4 ± 1.6	77.8 ± 0.8	75.0 ± 1.0	79.6 ± 0.8	79.2 ± 0.8	27.5 ± 0.1	39.6 ± 0.0	34.2 ± 0.0	68.3 ± 0.2	79.4 ± 0.1
Presto	ViT Nano	50.4 ± 0.3	–	–	52.5 ± 1.6	58.3 ± 1.4	59.1 ± 1.2	77.3 ± 1.3	71.2 ± 2.2	71.7 ± 1.5	14.0 ± 0.1	17.8 ± 0.2	20.0 ± 0.2	–	–
Tessera	–	–	–	–	–	–	72.1 ± 1.3	–	–	75.1 ± 3.7	–	–	37.7 ± 0.1	–	–
OlmoEarth	ViT Nano	62.8 ± 1.1	25.2 ± 0.5	22.7 ± 0.2	77.7 ± 0.9	57.8 ± 0.9	75.8 ± 0.7	78.7 ± 1.0	81.7 ± 1.0	81.4 ± 2.3	17.4 ± 0.6	33.5 ± 0.4	31.1 ± 0.5	55.2 ± 0.1	78.3 ± 0.1
OlmoEarth	ViT Tiny	63.2 ± 0.9	24.5 ± 0.3	23.0 ± 0.1	79.7 ± 0.4	58.3 ± 0.6	78.8 ± 1.0	77.5 ± 0.6	81.7 ± 1.1	79.7 ± 2.3	19.9 ± 1.0	39.0 ± 0.5	35.7 ± 0.8	58.2 ± 0.7	78.7 ± 0.2
OlmoEarth	ViT Base	68.7 ± 1.5	32.3 ± 0.6	28.7 ± 0.1	75.5 ± 1.5	59.8 ± 0.5	74.6 ± 1.1	74.1 ± 1.8	80.3 ± 0.9	81.2 ± 1.6	27.8 ± 0.9	49.3 ± 0.7	47.1 ± 1.0	67.2 ± 0.1	78.9 ± 0.1
OlmoEarth	ViT Large	67.5 ± 1.6	30.9 ± 0.4	28.4 ± 0.1	76.2 ± 1.6	59.4 ± 1.1	76.4 ± 2.1	74.7 ± 1.4	78.9 ± 0.7	79.2 ± 1.0	28.6 ± 0.7	51.0 ± 0.4	49.3 ± 0.6	66.5 ± 0.1	79.3 ± 0.2

Table 6. Linear Probing results across research benchmarks (mean ± half of the 95% CI over 10 seeds, a dash (-) indicates that the model does not support the input modality required by the task. Tasks where variability was less than 0.01 show up as 0.0.)

		m-bigearthnet	m-so2sat	m-brick-kiln	m-forestnet	m-eurosat	m-cashewplant	m-SA-crop-type	PASTIS	MADOS	Sen1Floods11
	Modalities	S2	S2	S2	L8	S2	S2	S2	S2	S2	S1
	Time series	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘
Model	Metric	μF1	Acc.	Acc.	Acc.	Acc.	mIOU	mIOU	mIOU	mIOU	mIOU
Anysat	ViT Base	68.7 ± 1.3	56.6 ± 4.9	98.7 ± 0.2	50.0 ± 3.5	95.5 ± 1.2	80.4 ± 1.5	34.2 ± 0.4	60.5 ± 0.8	65.8 ± 6.1	78.2 ± 1.6
Clay	ViT Large	65.5 ± 0.4	60.7 ± 1.5	98.6 ± 0.2	48.5 ± 5.8	95.8 ± 0.7	73.5 ± 1.4	33.2 ± 0.5	48.7 ± 0.7	68.9 ± 4.3	78.6 ± 0.2
CopernicusFM	ViT Base	71.1 ± 0.5	63.7 ± 6.9	98.0 ± 0.4	-	97.8 ± 1.5	78.9 ± 0.8	33.7 ± 0.8	54.6 ± 0.2	63.0 ± 6.8	78.2 ± 0.7
CROMA	ViT Base	69.4 ± 1.9	60.3 ± 2.6	98.7 ± 0.1	-	95.6 ± 0.6	46.7 ± 0.9	35.4 ± 1.2	56.4 ± 0.6	66.2 ± 4.5	79.5 ± 0.4
CROMA	ViT Large	71.5 ± 1.1	60.2 ± 3.8	98.0 ± 1.0	-	97.0 ± 1.3	47.6 ± 2.0	36.5 ± 1.1	58.1 ± 0.4	71.6 ± 7.5	79.2 ± 0.9
DINOv3 Sat	ViT Large	69.2 ± 1.8	62.6 ± 2.3	98.6 ± 0.8	**58.6 ± 4.9**	95.9 ± 1.7	80.8 ± 0.5	34.1 ± 1.0	42.8 ± 1.4	65.4 ± 1.8	-
Galileo	ViT Base	69.6 ± 0.9	64.9 ± 1.2	98.5 ± 0.4	-	97.7 ± 0.6	78.7 ± 0.2	36.1 ± 0.8	60.9 ± 0.6	72.3 ± 1.3	79.6 ± 0.4
Panopticon	ViT Base	69.5 ± 0.7	63.9 ± 4.4	**98.8 ± 1.2**	56.3 ± 0.8	**98.3 ± 0.1**	79.7 ± 0.2	33.7 ± 0.8	54.2 ± 0.4	74.1 ± 2.8	78.9 ± 0.7
Prithvi v2	ViT Huge	70.7 ± 0.3	63.9 ± 1.9	97.9 ± 0.7	53.1 ± 2.3	96.4 ± 1.3	**85.4 ± 18.7**	39.0 ± 0.7	58.7 ± 0.3	71.1 ± 5.0	-
Satlas	Swin Base	72.3 ± 0.8	65.1 ± 3.4	98.7 ± 0.2	55.7 ± 3.5	97.0 ± 0.4	76.7 ± 0.8	37.8 ± 0.3	57.0 ± 0.8	62.7 ± 4.8	78.6 ± 0.2
TerraMind	ViT Base	72.5 ± 0.3	64.4 ± 3.7	98.5 ± 0.4	-	97.4 ± 0.4	80.9 ± 0.3	39.4 ± 0.4	59.8 ± 0.3	71.5 ± 3.8	78.9 ± 1.5
TerraMind	ViT Large	**73.7 ± 1.4**	66.1 ± 1.5	98.3 ± 0.7	-	97.7 ± 0.1	81.3 ± 0.2	**41.2 ± 0.4**	60.7 ± 1.3	71.9 ± 1.5	**79.7 ± 0.7**
OlmoEarth (Random Init)	ViT Base	61.0 ± 0.2	47.4 ± 3.5	94.6 ± 0.6	43.3 ± 3.4	79.7 ± 1.7	36.9 ± 19.5	26.9 ± 2.0	43.3 ± 1.3	38.7 ± 20.5	76.5 ± 1.0
OlmoEarth	ViT Nano	66.5 ± 0.7	61.0 ± 1.5	98.1 ± 0.4	52.2 ± 4.0	95.1 ± 0.3	37.0 ± 9.9	34.9 ± 1.3	52.4 ± 2.1	59.5 ± 3.5	78.4 ± 0.8
OlmoEarth	ViT Tiny	69.8 ± 0.5	63.7 ± 1.3	98.2 ± 1.4	54.8 ± 6.6	96.8 ± 0.6	72.2 ± 0.9	38.4 ± 1.3	59.9 ± 0.8	68.9 ± 5.9	78.9 ± 1.7
OlmoEarth	ViT Base	72.0 ± 0.8	69.0 ± 1.2	98.5 ± 0.1	54.1 ± 2.0	98.0 ± 1.7	80.1 ± 0.7	39.9 ± 0.7	64.4 ± 0.3	77.8 ± 4.7	79.5 ± 0.8
OlmoEarth	ViT Large	72.5 ± 1.1	**69.9 ± 4.1**	98.5 ± 0.5	53.8 ± 2.8	**98.3 ± 0.5**	80.7 ± 0.7	40.6 ± 0.8	**65.7 ± 1.5**	**80.0 ± 7.2**	79.3 ± 0.9

Table 7. Fine-tuning results across research benchmarks (mean ± half of the 95% CI over three seeds, a dash (-) indicates that the model does not support the input modality required by the task. Bold numbers indicate the highest mean score for each task.)

While our model can benefit from labeled data we still see good performance with pure self-supervised training.

Building remote sensing foundation models necessitates some tradeoffs. While our final model is not the best in every metric it retains high performance across the board and has the best average score and lowest average per-task rank.

F. Comparison to AlphaEarth Foundations

The AlphaEarth foundation model [9] is comparable to OlmoEarth in that both draw on similar data sources and were designed to support similar downstream tasks. Rather than releasing the model, Google released only the global, annualized embeddings computed by AlphaEarth. We compare OlmoEarth both as a frozen feature extractor (where, like

	m-bigearthnet	m-so2sat	m-brick-kiln	m-forestnet	m-eurosat	BreizhCrops	PASTIS	PASTIS	MADOS	Sen1Floods11	Average	Average Rank
	S2 Acc.	S2 Acc.	S2 Acc.	L8 Acc.	S2 Acc.	S2 F1	S1 F1	S2 F1	S2 F1	S1 F1		
MAE	60.6	48.1	96.2	42.0	89.3	71.5	**31.1**	46.6	68.7	78.3	63.2	5.1
Only S2 Data	53.7	45.9	91.3	-	89.2	**71.7**	-	42.4	69.5	-	46.4	-
No Maps	59.5	58.6	95.2	**46.0**	92.6	71.4	29.1	48.0	70.2	77.9	64.9	4.7
No Agricultural Maps	60.9	66.5	94.3	**46.0**	93.9	71.4	29.0	48.5	71.4	78.8	66.1	3.6
Random Masking	60.7	**67.4**	94.7	43.5	91.8	70.3	24.7	51.1	71.9	77.8	65.4	4.7
No Inst. Contrastive Loss	60.5	65.6	93.6	44.9	93.6	70.2	28.5	51.4	72.1	78.4	65.9	4.7
Patch Disc Loss	62.0	62.1	**96.3**	44.8	94.0	70.3	29.6	50.0	**74.1**	**79.3**	66.2	3.0
Final Recipe	**62.3**	65.9	94.2	45.8	**94.6**	71.4	29.4	**52.2**	71.7	78.8	**66.6**	**2.9**

Table 8. Ablation experiment selectively removing components of OlmoEarth base model.

AlphaEarth, only embeddings are used) and as an end-to-end fine-tunable model.

It is expensive to export and download AlphaEarth embeddings from Google Earth Engine: our export jobs for 32×32 crops took 26 EECU-seconds on average, or $290 for a dataset with 100K crops. Thus, we were only able to evaluate AlphaEarth on five tasks: three classification tasks (Nandi, AWF, and Ecosystem), one per-pixel regression task (LFMC), and one segmentation task (Solar Farm).

Since the AlphaEarth model has not been released, we can't evaluate AlphaEarth under a finetuning regime. We assess the performance of the annualized AlphaEarth embeddings compared to the OlmoEarth embeddings from the ViT Base encoder using a simple KNN classifier. We use the timestep of AlphaEarth embeddings that has the highest overlap with the time range of the labels. To assess the benefits of more complex decoders, we use the partner task decoders described in Section 3.4, while sweeping over the input size (AlphaEarth embeddings already capture spatial context, so we find that a smaller input size performs better).

With a KNN-classifier, OlmoEarth outperforms AlphaEarth on the Nandi and AWF tasks, while AEF outperforms OlmoEarth on the Ecosystem mapping task. However, OlmoEarth benefits significantly from full fine-tuning, with the fine-tuned models outperforming the best possible with AlphaEarth on all five tasks. This underscores the value of an open model that makes per-task fine-tuning possible.

G. Patch Size Analysis for `m_cashew_plant`

We observe that for the `m_cashew_plant` evaluation task, larger patch sizes lead to better performance for models that support variable patch sizes, such as OlmoEarth and Galileo. Table 10 summarizes the linear probing and fine-tuning results for `m_cashew_plant` across different patch sizes.

This effect is unusual: a smaller patch size typically improves performance (e.g. Figure 4 of [53]). We hypothesize that this is due to the spatially coarse labels in the dataset, which are polygons instead of pixels (Figure 4).

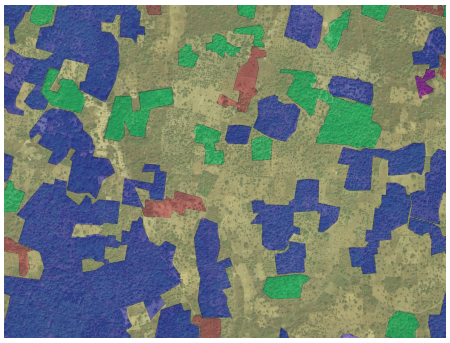

Figure 4. An example instance from the `m_cashew_plant` dataset: note the coarse, polygonal labels

H. OlmoEarth Platform

OlmoEarth Platform is an end-to-end solution that combines our foundation models with data management tools designed for organizations working on environmental challenges. The platform handles the complete workflow from satellite data collection through labeling, model fine-tuning, and inference, eliminating the need for organizations to manage GPU infrastructure or deep learning expertise. By making our models accessible, OlmoEarth Platform solves the last-mile problem of translating research into practical tools for applications including conservation, climate action, and food security.

		Nandi	AWF	Ecosystem	LFMC	Solar Farm
Model	Training	Acc.	Acc.	Acc.	L1	mIOU
AEF	kNN	55.6	81	60.6	-	-
AEF	Frozen + Decoder	66.0	75.9	61.2	23.1	77.5
AEF	Full Fine-tuning	Not Possible				
OlmoEarth	kNN	66.2	82	59.3	-	-
OlmoEarth	Frozen + Decoder	62.9	84.0	61.1	19.9	84.8
OlmoEarth	Full Fine-tuning	**82.2**	**86.0**	**62.4**	**17.9**	**86.7**

Table 9. Comparing AlphaEarth Foundation (AEF) embeddings with OlmoEarth ViT Base model using three different training strategies: kNN, frozen backbone + decoder, and decoder with full fine-tuning. For these evaluations, we use the "partner task" decoders described in Section 3.4.

Model	Patch 4×4		Patch 8×8		Patch 16×16	
	LP	FT	LP	FT	LP	FT
OlmoEarth-Base	27.7	71.9	27.9	76.2	32.3	79.8
Galileo-Base	24.3	73.0	25.6	76.9	28.9	78.8

Table 10. Performance (mIoU) comparison (LP = Linear Probing, FT = Fine-tuning) across patch sizes.

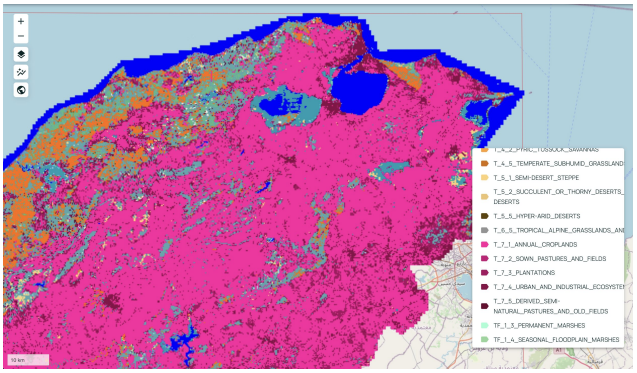

Figure 5. Results of a fine-tuned ecosystem classification model in the OlmoEarth Platform. Users can label data, fine-tune models, and run inference to generate maps all in the OlmoEarth Platform.

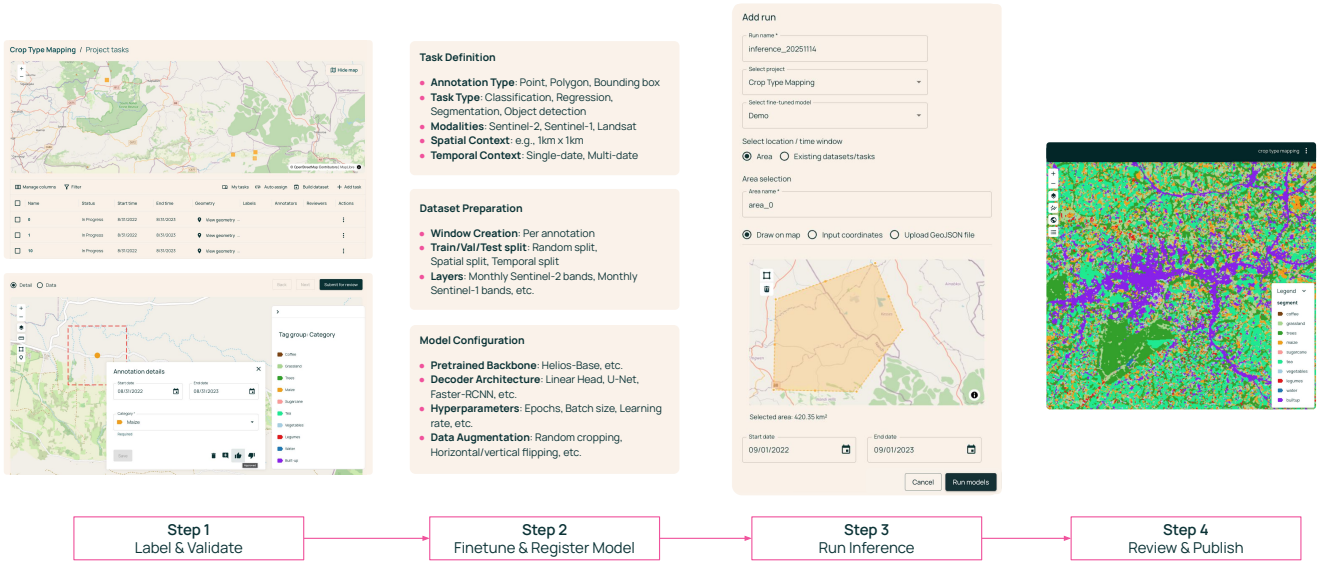

Figure 6. **OlmoEarth Platform: End-to-end workflow (using crop type mapping as an example).** The platform enables users to complete the full process from data labeling to map publishing: **Step 1:** Label and review annotations, **Step 2:** Fine-tune and register models for specific tasks, **Step 3:** Run inference on selected areas and time ranges, and **Step 4:** Review and publish the final maps.