# Pretrain Where? Investigating How Pretraining Data Diversity Impacts Geospatial Foundation Model Performance

Supplementary Material

## A. Additional Experimental Details

### A.1. Pretraining Dataset Construction

For pretraining SatMAE, we created seven new pretraining datasets. We maintained a dataset size comparable to the original FMoW-Sentinel dataset (approx. 700k samples). The primary aspects in which we diverged are as follows:

1. Scenes: FMoW-Sentinel is a 62-class scene classification dataset with images of urban structures. FMoW-Sentinel differs from the datasets we created as our images were chosen uniformly at random (UAR). Our pretraining captured random scenes without any structure.
2. Image size: FMoW-Sentinel, being a scene classification dataset, has images of random sizes, with variable heights and widths ranging from 50 to 500 pixels. But the SatMAE dataset pipelines randomly scale and crop images to a final size of $96 \times 96$. To limit the dataset size, we restricted the downloaded image resolution to $96 \times 96$. Despite the reduced resolution, we retained random scaling and cropping.

### A.2. Dataset Sampling Strategy

We now provide details on the data sampling procedure for our pretraining datasets. We created seven datasets: six continent-specific ones and one global dataset. We sampled data using the open-source software QGIS, which performs sampling operations efficiently. For the continent-specific datasets, we loaded a world continent map and sampled the required number of points using the *Random points in polygon* tool (Vector $>$ Research Tools). This sampled points uniformly at random, which satisfied the requirements of our use case.

### A.3. Dataset download

We downloaded Sentinel-2 images from Microsoft Planetary Compute, which provides free access to the full Sentinel-2 dataset. We used Python multiprocessing to download images centered on the sampled latitudes and longitudes. We limited the maximum cloud cover to 20% to obtain cloud-free samples. We initially sampled 900k points, retaining 700k for pretraining and 20k for validation to allow a buffer for download failures and missing images. We downloaded images from the year 2024. To further decrease dataset storage size, we normalised the images before saving, using the Sentinel-2 standard mean and standard deviation values from the SatMAE dataset loading pipeline available on GitHub. We made these datasets publicly available.

### A.4. Pretraining Hyperparameters

We performed pretraining using the original dataset pipelines provided on SatMAE's GitHub. We also used the same hyperparameters. Since SatMAE's original hyperparameters list is a bit confusing, we provide the hyperparameters we used here, for a H100-80GB GPU:

- batch size = 256
- accum iter = 16
- blr = 0.0001
- epochs = 50
- warmup epochs = 5
- input size = 96
- patch size = 8
- mask ratio = 0.75
- model = 'mae_vit_base_patch16'
- model type = 'group_c'

We pretrained for 50 epochs, as done for the original SatMAE pretraining, and noted convergence.

### A.5. Compute Infrastructure and Training Cost

All experiments were run on resources on ASU Sol supercomputer [2] and PSC bridges-2 [1] resources. We utilized two distinct compute resources: NVIDIA A100-40GB (ASU Sol) and NVIDIA H100-80GB GPUs (PSC Bridges-2). We distributed the workload for pretraining and downstream tasks across both environments. The estimated runtimes for each setup are detailed below.

**NVIDIA A100 (40GB) Benchmarks**

- Pretraining: $\approx$ 48 hours
- FMoW Evaluation: $\approx$ 60 minutes
- Mosaiks Evaluation: $\approx$ 20 minutes
- ForTy Evaluation: $\approx$ 20 minutes
- GeoBench Evaluation: $\approx$ 300 minutes

**Total A100 Cost:** Pretraining ($48\text{h} \times 10$) + Downstream ($400\text{min} \times 11 \times 7 \times 5$).

**NVIDIA H100 (80GB) Benchmarks**

- Pretraining: $\approx$ 10 hours
- FMoW Evaluation: $\approx$ 15 minutes
- Mosaiks Evaluation: $\approx$ 5 minutes
- ForTy Evaluation: $\approx$ 5 minutes
- GeoBench Evaluation: $\approx$ 70 minutes

**Total H100 Cost:** Pretraining ($10\text{h} \times 10$) + Downstream ($100\text{min} \times 11 \times 7 \times 5$).

Since we utilised both server types, our true computational cost is split between the two.

# B. Model and Training Details

## B.1. SatMAE Architecture choices

We utilised the ViT-B architecture for our experiments to optimise computational resources. As demonstrated in the SatMAE paper, the performance gap between ViT-B and ViT-L was minimal. We observed similar trends in our preliminary experiments; therefore, we proceeded with ViT-B to reduce training costs. We used the standard architecture without any modifications.

## B.2. Pretraining Objective and Optimisation

We provided all the pretraining hyperparameters in A.4. We did not make any other changes apart from moving data normalisation to the download stage for faster dataset loading. The pretraining objective remained unchanged.

## B.3. Linear Probing Protocol

We added linear classifiers for each downstream task, as the original SatMAE code was limited to full finetuning for classification tasks. For all downstream tasks, we fixed the training duration to 50 epochs, as we observed convergence within this timeframe. We used a cosine decay schedule with 5 warmup epochs and a batch size of 512. We swept over a large set of learning rates specific to each downstream task; we ensured the sweep is exhaustive by verifying that the optimal learning rate is bounded by lower-performing rates on both sides. To satisfy this criterion, we swept across a range of $\{1, 3, 5, 8\} \times \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ to identify optimal values.

# C. Downstream Datasets and Evaluation

## C.1. FMoW-Sentinel Dataset Details and Splits

We used the publicly available FMoW-Sentinel dataset. To create continent-specific subsets, we split the global dataset using QGIS. We imported the world continents map and the FMoW-Sentinel image polygons, then computed the intersection of the two vector layers. To create smaller subsets capped at $\approx$ 5k samples, we restricted the dataset to the 20 most frequent classes per continent. To achieve a 70:15:15 (train:val:test) split, we limited sampling to $\approx$ 250 images per class (175 for train, 38 for validation, and 38 for test). This resulted in a total dataset size of $20 \times 251 \approx$ 5k samples.

For the global subset, we sampled from the continent-specific subsets. From each subset, we selected 44 samples per class: 30 for training, 7 for validation, and 7 for testing.

The total dataset size was $(30 + 7 + 7) \times 20 \times 6 \approx$ 5k samples.

To generate 5 data seeds, we created 5 distinct versions of all subsets by changing the random seed during the sampling procedure.

## C.2. MOSAIKS Population Density Dataset

We constructed this particular dataset. We took the labels with their latitude and longitude coordinates from the MOSAIKS paper. We downloaded the corresponding Sentinel-2 images similar to how we did for our custom pretraining datasets mentioned in A.3. Similar to FMoW-Sentinel, we divided the full dataset into per-continent partitions and then sampled 5k points. We divided the points into 70:15:15-sized splits. 5 data seeds were created as done for FMoW-Sentinel.

## C.3. ForTy Segmentation Dataset

We used the publicly available ForTy dataset in tfrecords format. We extracted point coordinates from the tfrecords and followed procedures similar to MOSAIKS for creating subsets.

## C.4. GEO-Bench

We did not alter anything in the GEO-Bench datasets. We used the 6 datasets that contained data from Sentinel-2 and used the 1.00x partitions, which have 100% data. We did not work with the data seeds given for GEO-Bench. We now give details for individual GEO-Bench tasks that we worked on:

- **m-eurosat:** A 4k-sized, 10-class scene classification dataset with samples from Europe only.
- **m-bigearthnet:** A 22k-sized, 43-class land cover classification dataset with samples from Europe only.
- **m-brick-kiln:** A 17k-sized, 2-class brick kiln classification dataset with samples from Bangladesh, Asia only.
- **m-so2sat:** A 21k-sized, 17-class land cover classification with global samples.
- **m-cashew-plantation:** A 1.8k-sized, 7-class cashew plantation identification segmentation task with samples from Benin, Africa.
- **m-sa-crop-type:** A 5k-sized, 10-class crop type classification dataset with samples from South Africa, Africa.

## C.5. Evaluation Metrics

We used Accuracy, R2 score, and F1Score for FMoW-Sentinel, MOSAIKS, and Forty, respectively. For GEO-Bench, we worked with the standard metrics that came with each task, i.e., Jaccard for m-SA-crop-type and m-cashew-plantation, F1Score for m-bigearthnet, and Accuracy for the rest.

# D. Additional Results

## D.1. Per-Continent Downstream Results

We present per-continent downstream results drawn as heatmaps for MOSAIKS and ForTy tasks here:

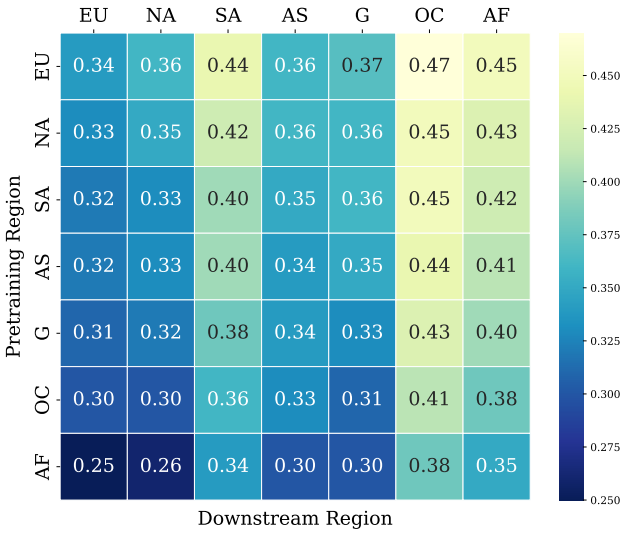


Figure 1. Performance comparison on ForTy global subsets across pretraining data schemes. In-distribution sampling is not always optimal.

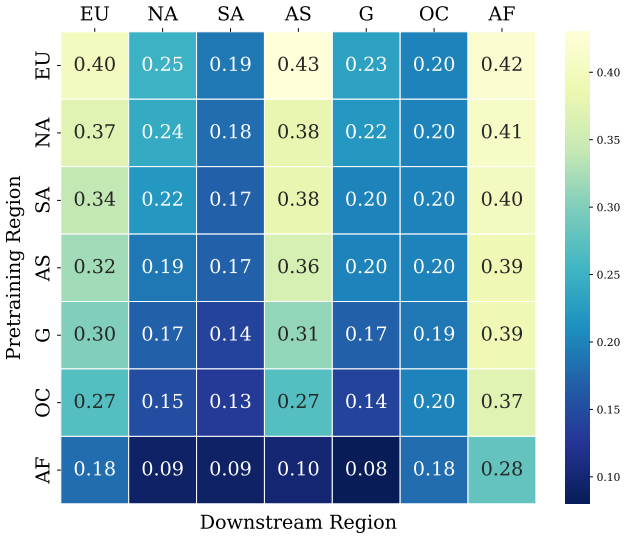


Figure 2. Performance comparison on MOSAIKS population density global subsets across pretraining data schemes. In-distribution sampling is not always optimal.

# E. Dataset Diversity Analysis

## E.1. Other pretrainings

For the diversity analyses, we also worked with 3 additional published datasets, FMoW-Sentinel, SSL4Eo and SSL4Eco.

- FMoW-Sentinel: FMoW-Sentinel is a 700k-sized 62-class scene classification dataset, originally used for unsupervised pretraining by SatMAE. The distribution is biased towards the Global North. The spatial distribution of FMoW is [0.21, 0.09, 0.35, 0.23, 0.08, 0.02] in terms of sample distribution for continents Asia, Africa, Europe, North America, South America, and Oceania.
- SSL4Eco: is a 1 M-sized global pretraining dataset. SSL4Eco is seasonal, i.e., captures 4 images from 4 seasons for 250k locations to reach 1M. The authors of SSL4Eco also distinguish SSL4Eco as a dataset with samples from all Copernicus landcover classes. For sampling, the dataset followed a uniform grid strategy from MajorTOM, but with 23km spacing between any two points. The spatial distribution is [0.32, 0.21, 0.07, 0.17, 0.12, 0.05]
- SSL4Eo: SSL4Eo is also sized at 1M, with sampling focused around city centers. The SSL4Eo authors chose city centers and then sampled around them in a radius of 50km using Gaussian sampling. They also ensured removal of overlapping samples.

## E.2. Geographic Class Definitions (Continents, Biomes, Landcover)

For diversity calculations, we use three approaches - continent-based, biome-based, and landcover-based. Notice that in each of these approaches, the dataset can be partitioned into multiple groups/classes. Specifically:

- Continents: have 6 groups, namely Asia, Africa, Europe, North America, South America, Oceania.
- Biomes: According to the RESOLVE biome map, there are 15 biomes namely (1) Deserts & Xeric Shrublands, (2) Tropical & Subtropical Grasslands, Savannas & Shrublands, (3) Boreal Forests/Taiga, (4) Tundra, (5) Tropical & Subtropical Moist Broadleaf Forests, (6) Temperate Broadleaf & Mixed Forests, (7) Temperate Grasslands, Savannas & Shrublands, (8) Mediterranean Forests, Woodlands & Scrub, (9) Montane Grasslands & Shrublands, (10) Tropical & Subtropical Dry Broadleaf Forests, (11) Temperate Conifer Forests, (12) Flooded Grasslands & Savannas, (13) Tropical & Subtropical Coniferous Forests, (14) Mangroves, (15)rock and ice.
- Landcover: According to ESA Worldcover 2021 v200, landcover has 11 classes namely (1) Tree cover, (2) Shrubland, (3) Grassland, (4) Cropland, (5) Built-up, (6)Bare / sparse vegetation, (7) Snow and ice, (8) Permanent water bodies, (9) Herbaceous wetland, (10) Mangroves, (11) Moss and lichen

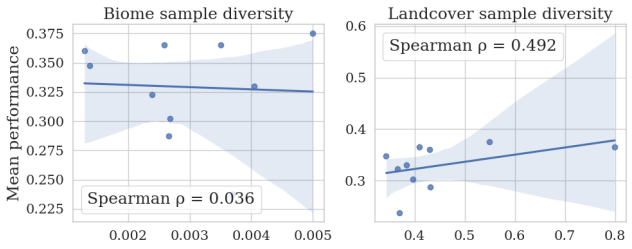


Figure 3. Correlation plots between mean model performance and diversity measures: sample biome and sample landcover diversity.

### E.3. More diversity measures

We now look at biome and landcover diversity from the sample entropy perspective similar to how the spectral diversity is calculated. We calculate entropy for each sample and then take the mean over the dataset. We found that sample biome diversity (shown in Figure 3, left) had a lower correlation with downstream performance as compared to biome diversity. On the other hand, sample landcover diversity (shown in Figure 3, right) has a higher correlation than lancover diversity.

Note that the definitions of biome and landcover diversity used in the main paper align more with biome/landcover based stratified sampling or sampling to ensure data from all biomes/landcovers for example, MMEarth [3], SSL4Eco [4], Presto [5], etc. While the definitions defined here are related to diversity ensured within a sample, which is implicitely done in sampling approach of Galileo [6].

### E.4. Per-band spectral diversity

Since spectral diversity is calculated by first calculating entropy for a sample's band and then averaging over bands, we also find a dataset's diversity in terms of its individual bands. For this, we calculate a list of entropies for each sample corresponding to the bands, and then average them across the dataset. We show the values for the One-hot-Europe dataset in Table 1

We note that the RGB bands B4, B3 and B2 show the least entropy whereas bands like B6, B7, B8, B8A (Red Edge bands normally used for vegetation related tasks) have higher entropy.

Table 1. Per-band spectral diversity for One-hot-Europe pretraining dataset.

| Band | Spectral entropy |
|---|---|
| B2 (blue) | 2.20 |
| B3 (green) | 2.38 |
| B4 (red) | 2.22 |
| B5 (red edge) | 2.46 |
| B6 | 2.60 |
| B7 | 2.63 |
| B8 (NIR) | 2.56 |
| B8A | 2.64 |
| B11 (SWIR 1) | 2.54 |
| B12 (SWIR 2) | 2.37 |

## References


[1] S. T. Brown, P. Buitrago, E. Hanna, S. Sanielevici, R. Scibek, and N. A. Nystrom. Bridges-2: A platform for rapidly-evolving and data intensive research. In *Practice and Experience in Advanced Research Computing*, pages 1–4, 2021. 1

[2] Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, Will Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobwala, Thirugnanam Jagadeesan, Praful Barghav, Torey Battelle, Rebecca Belshe, Debra McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Sean Dudley, Dhruvil Shah, Gil Speyer, and Jason Yalim. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing*, pages 296–301, New York, NY, USA, 2023. Association for Computing Machinery. 1

[3] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024. 4

[4] Elena Plekhanova, Damien Robert, Johannes Dollinger, Emilia Arens, Philipp Brun, Jan Dirk Wegner, and Niklaus E. Zimmermann. Ssl4eco: A global seasonal dataset for geospatial foundation models in ecology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2428–2439, 2025. 4

[5] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah R Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. 4

[6] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global amp; local features of many remote sensing modalities. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 60280–60300. PMLR, 2025. 4