How Does the Spatial Distribution of Pre-training Data Affect Geospatial Foundation Models?

Mirali Purohit*[1], Gedeon Muhawenayo*[1], Esther Rolf[2], Hannah Kerner[1]

[1]Arizona State University
[2]University of Colorado Boulder

Abstract

Foundation models have made rapid advances in many domains including Earth observation, where Geospatial Foundation Models (GFMs) can help address global challenges such as climate change, agriculture, and disaster response. Previous work on GFMs focused on tailoring model architecture and pre-text tasks, and did not investigate the impact of pre-training data selection on model performance. However, recent works from other domains show that the pre-training data distribution is an important factor influencing the performance of the foundation models. With this motivation, our research explores how the *geographic* distribution of pre-training data affects the performance of GFMs. We evaluated several pre-training data distributions by sampling different compositions from a global data pool. Our experiments with two GFMs on downstream tasks indicate that balanced and globally representative data compositions often outperform region-specific sampling, highlighting the importance of diversity and global coverage in pre-training data. Our results suggest that the most appropriate data sampling technique may depend on the specific GFM architecture. These findings will support the development of robust GFMs by incorporating quality pre-training data distributions, ultimately improving machine learning solutions for Earth observation.

1 Introduction

Foundation models have proven to be powerful tools across a wide variety of fields. These models have shown remarkable potential by achieving promising results while reducing the dependencies on labeled data. Researchers are increasingly adapting foundation models in specialized domains, such as health, law, physics, and geoscience by leveraging their ability to generalize from large datasets (Colombo et al. 2024; Parmar et al. 2022; Cong et al. 2022).

In recent years, Geospatial Foundation Models (GFMs) pre-trained on vast quantities of satellite data have been developed for Earth observation (EO) applications. These models can play a critical role in improving the performance of tasks related to climate change, agriculture, oceanography, forests, disaster response, and more (Jakubik et al. 2023; Tseng et al. 2023; Klemmer et al. 2023).

There is a general perception in AI that more data will help to improve the model's performance (Sun et al. 2017; Shi et al. 2024). Hence, most current efforts to build foundation models usually pre-train on as much data as available or computationally feasible. However, recent studies show that dataset characteristics like diversity, spatial distribution, and representivity can be more important than data quantity to build robust ML models (Nguyen et al. 2022; Cole et al. 2022). This suggests that understanding the effects of pre-training data distribution on model performance is crucial.

Although some studies have explored the relation between pre-training data distribution and downstream task performance in computer vision (CV) for natural images (Al Kader Hammoud et al. 2024; Fang et al. 2022; Longpre et al. 2024), there is a lack of studies of this type for GFMs. To build GFMs, researchers use large remote sensing datasets collected by satellites and aerial platforms. Given the massive data volume available from EO satellites (e.g., from the Sentinel-2 satellites), it is essential to select pre-training data in a way that benefits model performance but also reduces the computational cost – e.g., by focusing on data quality instead of quantity (Rolf et al. 2021).

Unlike standard CV datasets, which are generally static, geospatial datasets are dynamic and available from multiple time periods. Satellite data offers the unique advantage of global coverage, allowing data to be sampled from any location on Earth. However, we don't know the answer to some critical questions, such as "Is it crucial to have data from all continents? Is it crucial to have data evenly represented across biomes?"

As a result, existing GFMs employ different strategies to sample pre-training data, showing a *lack of consensus on a unified approach to sample pre-training data* (details in §2). Roscher et al. (2024) showed that most of the modeling focus in the geospatial domain is currently paid to the pre-training approach or model architecture. At the same time, *the composition of data chosen for pre-training is relatively understudied*, due to the assumption that training data size is the dominant factor determining performance. In particular, Manas et al. (2021) is the only study that compares model performance with multiple data sampling strategies.

Motivated by these limitations, we investigated how different GFMs perform on downstream tasks when different sampling techniques are used to curate pre-training data

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

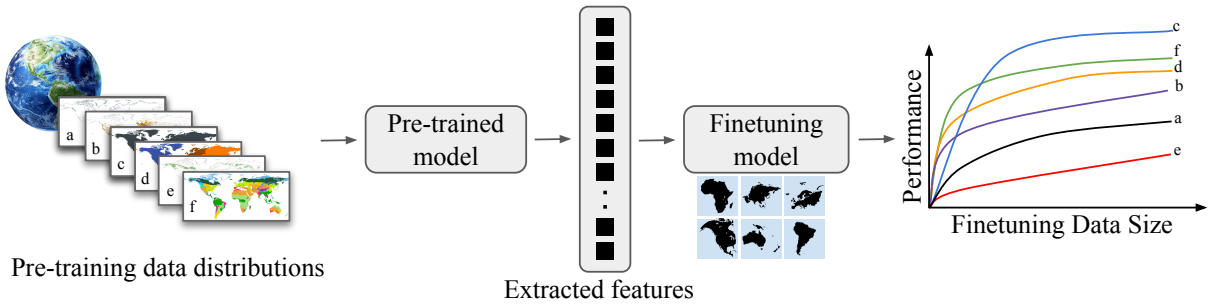


Figure 1: Our experimental pipeline to measure the impact of pre-training data distributions on downstream task performance.

(pipeline shown in Figure 1). We conducted experiments on two GFMs by pre-training models on five different data sampling strategies, referred as *data compositions*. These strategies include random, stratified, and clustered sampling techniques (see §3.3). We refer to sampling techniques that ensure balanced or global data representation as balanced sampling techniques, and those grouping samples by spatial or environmental characteristics as clustered sampling. We pre-trained the models on five different data compositions and investigated how they affect the downstream task performance disaggregated by continent (Figure 1). We compared these results with a baseline of no pre-training. Our main contributions are:

- We investigated the effect of different pre-training data compositions on downstream tasks for GFMs. To the best of the authors' knowledge, this is the first study to explore the impact of pre-training data distribution on downstream task performance for GFMs.
- Our experimental results show that all balanced sampling techniques yield approximately equal performance on downstream tasks, and usually outperform models with clustered pre-training data or no pre-training.

Overall, our observations highlight the importance of balanced and representative data sampling strategies in pre-training GFMs. We believe our findings will contribute to the development of robust GFMs by incorporating quality pre-training data that benefit downstream task performance.

2 Related Work

The relationship between pre-training data distribution and its impact on downstream tasks is well-documented in the CV literature (detailed discussion in Appendix A). However, this relationship for GFMs has received limited attention.

Previous GFMs used heuristic measures of diversity to decide where to sample pre-training data, with large differences from model to model. Presto developed a stratified random sampling strategy across hemispheres, ecological biomes, and land cover types (Tseng et al. 2023). MMEarth sampled pre-training data uniformly across biomes and from five years (Nedungadi et al. 2024). SatCLIP used a global uniform at random sampling (Klemmer et al. 2023). SeCo, SSL4EO, and CROMA used pre-training data within a range of 50km around 10,000 largest cities globally, assuming data will be more diverse around cities (Manas et al. 2021; Wang et al. 2023; Fuller, Millard, and Green 2024). Prithvi used pre-training data only from the contiguous United States (Jakubik et al. 2023). SatMAE, ScaleMAE, and SpectralGPT pre-trained on existing datasets fMoW (Christie et al. 2018), which is biased towards the Global North, and BigEarthNet (Sumbul et al. 2019), which includes only Europe (Cong et al. 2022; Reed et al. 2023; Hong et al. 2024).

Previous work chose different pre-training data sampling techniques due to differing assumptions about which strategy yields a diverse data distribution. The impact of different sampling methods on downstream task performance remains unclear, as there has been no systematic comparison across different pre-training sampling techniques. In this research, we aim to fill this gap by providing empirical insights and guidance for future GFM development.

3 Experimental Setup

3.1 Task Formulation

Our goal is to investigate how different pre-training data compositions affect the performance of GFMs on downstream tasks. We performed experiments with two GFMs (explained in §3.2), each pre-trained on five different data compositions created by various sampling techniques from globally available data, and compared with a baseline of no pre-training (details in §3.3).

For finetuning, we created subsets of downstream data samples based on their continent of origin. We split each continent subset into training, validation, and testing sets. We then finetuned each pre-trained model on task-specific training data from each continent (Africa, Asia, Europe, North America, Oceania, and South America) to investigate each model's performance in few-shot settings. This continent-wise approach allowed us to assess the model's effectiveness with region-specific tasks, which capture unique geographic and continental variations. We finetuned and evaluated each GFM on a specific downstream task explained in §3.2.

Since collecting labeled data for geospatial applications can be expensive (Rolf et al. 2024), there is a need for models that can achieve good performance in a few-shot setting. We finetuned all downstream tasks on a few-shot learning setting by randomly selecting a total of $n = 100$ samples

from each continent's training set (Schick and Schütze 2021; Pecher, Srba, and Bielikova 2024).

To improve the reliability of the results for few-shot finetuning, we repeated each finetuning experiment 50 times with a different random seed for sampling training data. We reported the average and standard deviation across these iterations for each continent. We finetuned and evaluated each task using both parametric (Multi-Layer Perceptron (MLP) or logistic regression) and non-parametric (Random Forest and K-Nearest Neighbors (KNN)) models. Details of the experimental configuration are provided in Appendix C.

3.2 Geospatial Foundation Models and Downstream Tasks

Presto Presto is a lightweight pixel timeseries foundation model that aims to capture temporal patterns and spatial features efficiently (Tseng et al. 2023). It is one of the few GFMs that uses time series satellite data to capture seasonal patterns. Presto uses various spectral bands from different satellite sensors, topographic data, location coordinates, etc.

We evaluated Presto using the CropHarvest agriculture downstream task (Tseng et al. 2021). We chose CropHarvest because it was used for evaluation in the original Presto paper and has global coverage. CropHarvest is a binary crop vs. non-crop classification task. We created a custom task from CropHarvest instead of using the pre-defined benchmark tasks to provide continent-wise subsets for finetuning (explained in the Appendix C).

SatCLIP SatCLIP is a location encoder-based foundation model (Klemmer et al. 2023), with a strong capability to learn implicit representations of locations. As a model optimized for extracting location-based features from satellite imagery, it effectively captures geographic and environmental characteristics. Being a location encoder, the input of each task consists of raw latitude and longitude coordinates.

We evaluated SatCLIP with the EcoRegions downstream task (Dinerstein et al. 2017) which is a multi-class classification of 14 classes. For this task, the model classifies ecological zones into predefined biome categories. This task assesses the model's understanding of geographical and ecological patterns, which include climate, soil type, vegetation, and biodiversity. We chose the EcoRegions task because it was used for evaluation in the original SatCLIP paper and has global coverage.

3.3 Pre-training Data Compositions

To vary the geographic distribution of pre-training data, we implemented several pre-training data sampling strategies, ranging from uniform random sampling to targeted regional geographies. Each technique was selected to capture different geographic data variability that could impact GFM performance. We experimented with five different data compositions, and zero pre-training as a baseline, detailed below:

- Zero pre-training (Zero): To establish a baseline for assessing the impact of pre-training data, we employed zero pre-training in which the GFM is finetuned from random weights. This measures the model's performance solely based on its architecture and finetuning data.
- Uniform at Random (UAR): To capture global representation, we uniformly and randomly sampled data with equal likelihood from the Earth's landmass. This sampling approach was used in Klemmer et al. (2023) to minimize bias toward any specific region or environment.
- Stratified Continent: We sampled pre-training data stratified by continents to achieve balanced representation across continental boundaries (Esri 2023). Within each stratum, we used uniform at random sampling to select points. We excluded Antarctica due to its limited land use and diversity.
- Stratified Biome: To capture environmental biodiversity and landcover, we uniformly sampled within each biome stratum, drawing an equal number of samples from the ecological zones created by Dinerstein et al. (2017).
- Natural Forest: To evaluate the model's capability in a strongly biased data context, we sampled all data exclusively from global intact forest cover (Potapov et al. 2017). This simulates cases where researchers may develop domain-specific foundation models, e.g., for forestry-related tasks.
- World Cities: In this population-centric sampling strategy, we sampled randomly within a 50 km radius around the world's 10,000 most populated cities (SimpleMaps 2024). This emulates the approach used by Manas et al. (2021); Wang et al. (2023); Fuller, Millard, and Green (2024) which sampled using a Gaussian distribution around the same cities, assuming land cover/land use is most diverse near cities.

We created pre-training datasets for SatCLIP using the Microsoft Planetary Computer data catalog (Source et al. 2022). For Presto, we drew samples from the global pool of $\sim$ 22 million samples provided by (Tseng et al. 2023). We provided further details on the composition of each pre-training dataset, including the number of samples and visualizations of their geographic distributions, in Appendix B.

4 Results and Analysis

Table 1 shows the F1 score for the CropHarvest and EcoRegions tasks using Random Forest (results of other finetuning models are included in Appendix D). We summarize our results in response to specific research questions below.

How do different data compositions affect the performance of downstream tasks? Table 1 shows that models pre-trained with any data composition generally outperform models with zero pre-training. Overall, all balanced data sampling techniques (i.e., UAR, stratified continent/biome) outperform clustered techniques, i.e., Natural Forest for CropHarvest and both Natural Forest and World Cities for EcoRegions. Notably, performance is approximately consistent among balanced techniques, suggesting that *among sampling strategies that ensure global representation, there is less difference in downstream task performance*. This finding can be further strengthened in future work by evaluating more balanced and clustered sampling techniques and additional downstream tasks. In CropHarvest, for South America and Oceania, Natural Forest performs on par with other data

Table 1: Average F1 score and standard deviation (over 50 runs) for the CropHarvest and EcoRegions when Presto and SatCLIP are finetuned using Random Forest (RF) on 100 training samples, respectively. Here, highlighted results indicate scores that are at least 2% lower than UAR. (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America)

Data Composition	CropHarvest						EcoRegion					
	AF	AS	EU	NA	OC	SA	AF	AS	EU	NA	OC	SA
Zero	0.68 ± 0.02	0.76 ± 0.03	0.74 ± 0.02	0.79 ± 0.02	0.73 ± 0.03	0.75 ± 0.02	0.66 ± 0.02	0.47 ± 0.03	0.67 ± 0.02	0.61 ± 0.03	0.73 ± 0.05	0.58 ± 0.04
UAR	0.71 ± 0.02	0.78 ± 0.03	0.76 ± 0.02	0.80 ± 0.03	0.75 ± 0.03	0.79 ± 0.02	0.78 ± 0.01	0.68 ± 0.03	0.81 ± 0.02	0.77 ± 0.02	0.84 ± 0.02	0.76 ± 0.03
Stratified Continent	0.71 ± 0.02	0.78 ± 0.03	0.76 ± 0.03	0.81 ± 0.03	0.75 ± 0.02	0.80 ± 0.02	0.75 ± 0.02	0.66 ± 0.02	0.81 ± 0.02	0.76 ± 0.02	0.83 ± 0.03	0.75 ± 0.02
Stratified Biome	0.71 ± 0.02	0.78 ± 0.03	0.75 ± 0.02	0.81 ± 0.02	0.76 ± 0.03	0.80 ± 0.02	0.77 ± 0.02	0.66 ± 0.02	0.82 ± 0.02	0.75 ± 0.04	0.83 ± 0.02	0.75 ± 0.03
Natural Forest	0.67 ± 0.02	0.75 ± 0.03	0.73 ± 0.02	0.77 ± 0.03	0.75 ± 0.03	0.80 ± 0.03	0.59 ± 0.03	0.42 ± 0.02	0.65 ± 0.04	0.58 ± 0.03	0.76 ± 0.03	0.61 ± 0.04
World Cities	0.72 ± 0.02	0.78 ± 0.03	0.76 ± 0.03	0.81 ± 0.03	0.75 ± 0.02	0.79 ± 0.02	0.72 ± 0.02	0.57 ± 0.03	0.78 ± 0.02	0.61 ± 0.02	0.77 ± 0.03	0.72 ± 0.03

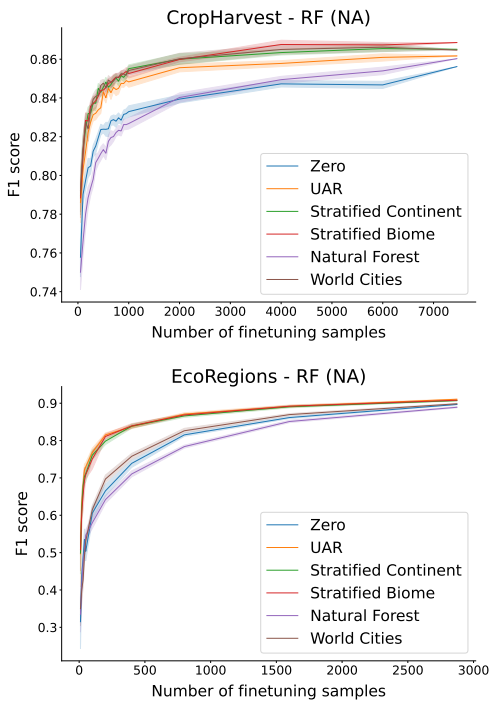


Figure 2: Performance Comparison across different data compositions for North America (NA).

compositions, likely because Natural Forest covers a large portion of these continents.

Do all data compositions show similar trends across different GFMs? In other words, can we expect that a specific data composition performs similarly across different models compared to other compositions? Our results indicate that the relative order of performance for different data compositions varies for each model. Table 1 shows that clustered data compositions, i.e., Natural Forest and World Cities, perform poorly in EcoRegions, however, World Cities shows an equal level of performance with other data compositions in CropHarvest. The reason behind this discrepancy could be the architectural design of the model. SatCLIP, which is a location-encoder model, struggles to generalize when pre-training data lacks global representation. However, Presto is a pixel timeseries model (which does not rely solely on location) and is likely more capable of generalizing on World Cities – which is less clustered than Natural Forest. This hypothesis can be investigated in future work by performing experiments with additional GFMs based on location-encoding and pixel timeseries.

Do performance differences between data compositions disappear when models are finetuned with an increasing amount of data? Our findings indicate that pre-training data compositions impact how models perform with different amounts of finetuning data. Balanced data compositions outperform clustered data sampling in few-shot scenarios. However, when finetuning is conducted on an increasing amount of data (thousands of samples), the performance differences between various data compositions drop significantly. As shown in Figure 2 for North America, lines representing different data compositions converge as the number of finetuning samples increases (results of other continents are presented in the Appendix). Similar findings were observed across nine downstream tasks with ImageNet or LAION pre-training Entezari et al. (2023).

5 Conclusions

As the development of GFMs advances, focusing on quality of pre-training data is essential, yet the impact of pre-training data distribution on downstream performance remains underexplored. To address this gap, our study is the first to investigate how pre-training data distribution can affect downstream task performance for GFMs. We conducted experiments on two GFMs by pre-training with different data compositions (including random, stratified, and clustered sampling) and evaluating in few-shot settings.

Experimental results indicate that balanced and global representative sampling techniques generally outperform clustered or region-specific compositions, highlighting the importance of globally diverse pre-training data for GFMs. Results also show that the effects of data sampling techniques can vary depending on model architecture. Moreover, balanced sampling techniques exhibit nearly equal performance, suggesting that the choice of specific sampling techniques is insignificant, as long as the composition ensures balanced samples across the globe.

Our results provide preliminary insights into how the geographic distribution of pre-training data can influence downstream task performance. However, further exploration can be done to deepen our understanding by adding more GFMs and data sampling strategies. Detailed limitations and directions for future work are provided in the Appendix E.

References

Abbasi, R.; Rohban, M. H.; and Baghshah, M. S. 2024. Deciphering the Role of Representation Disentanglement: Investigating Compositional Generalization in CLIP Models. In *European Conference on Computer Vision*. Springer.

Al Kader Hammoud, H. A.; Das, T.; Pizzati, F.; Torr, P. H.; Bibi, A.; and Ghanem, B. 2024. On Pretraining Data Diversity for Self-Supervised Learning. In *European Conference on Computer Vision*, 54–71. Springer.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional Map of the World. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180.

Cole, E.; Yang, X.; Wilber, K.; Mac Aodha, O.; and Belongie, S. 2022. When Does Contrastive Visual Representation Learning Work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14755–14764.

Colombo, P.; Pires, T. P.; Boudiaf, M.; Culver, D.; Melo, R.; Corro, C.; Martins, A. F.; Esposito, F.; Raposo, V. L.; Morgado, S.; et al. 2024. Saullm-7b: A Pioneering Large Language Model for Law. *arXiv preprint arXiv:2403.03883*.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-spectral Satellite Imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Dinerstein, E.; Olson, D.; Joshi, A.; Vynne, C.; Burgess, N.; Wikramanayake, E.; Hahn, N.; Palminteri, S.; Hedao, P.; Noss, R.; Hansen, M.; Locke, H.; Ellis, E.; Jones, B.; Barber, C.; Hayes, R.; Kormos, C.; Martin, V.; Crist, E.; and Saleem, M. 2017. An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, 1: 1–12.

Entezari, R.; Wortsman, M.; Saukh, O.; Shariatnia, M. M.; Sedghi, H.; and Schmidt, L. 2023. The Role of Pre-training Data in Transfer Learning. *arXiv preprint arXiv:2302.13602*.

Esri. 2023. World Continents. Accessed January 27, 2024.

Fang, A.; Ilharco, G.; Wortsman, M.; Wan, Y.; Shankar, V.; Dave, A.; and Schmidt, L. 2022. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In *International Conference on Machine Learning*, 6216–6234. PMLR.

Fuller, A.; Millard, K.; and Green, J. 2024. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36.

Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. 2024. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. *arXiv preprint arXiv:2310.18660*.

Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; and Rußwurm, M. 2023. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. *arXiv preprint arXiv:2311.17179*.

Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; and Ippolito, D. 2024. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3245–3276. Mexico City, Mexico: Association for Computational Linguistics.

Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; and Rodriguez, P. 2021. Seasonal Contrast: Unsupervised Pre-training from uUncurated Remote Sensing Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9414–9423.

Nedungadi, V.; Kariryaa, A.; Oehmcke, S.; Belongie, S.; Igel, C.; and Lang, N. 2024. MMEarth: Exploring Multi-Modal Pretext Tasks For Geospatial Representation Learning. In *European Conference on Computer Vision*. Springer.

Nguyen, T.; Ilharco, G.; Wortsman, M.; Oh, S.; and Schmidt, L. 2022. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. *Advances in Neural Information Processing Systems*, 35: 21455–21469.

Parmar, M.; Mishra, S.; Purohit, M.; Luo, M.; Mohammad, M.; and Baral, C. 2022. In-BoXBART: Get Instructions into Biomedical Multi-Task Learning. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 112–128. Seattle, United States: Association for Computational Linguistics.

Pecher, B.; Srba, I.; and Bielikova, M. 2024. Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance. *arXiv preprint arXiv:2402.12819*.

Potapov, P.; Hansen, M. C.; Laestadius, L.; Turubanova, S.; Yaroshenko, A.; Thies, C.; Smith, W.; Zhuravleva, I.; Komarova, A.; Minnemeyer, S.; et al. 2017. The Last Frontiers of Wilderness: Tracking Loss of Intact Forest Landscapes from 2000 to 2013. *Science advances*, 3(1): e1600821.

Ramanujan, V.; Nguyen, T.; Oh, S.; Farhadi, A.; and Schmidt, L. 2024. On the Connection between Pre-training Data Diversity and Fine-tuning Robustness. *Advances in Neural Information Processing Systems*, 36.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-MAE: A Scale-aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Rolf, E.; Klemmer, K.; Robinson, C.; and Kerner, H. 2024. Mission Critical–Satellite Data is a Distinct Modality in Machine Learning. In *International Conference on Machine Learning*. PMLR.

Rolf, E.; Worledge, T. T.; Recht, B.; and Jordan, M. 2021. Representation matters: Assessing the Importance of Subgroup Allocations in Training Data. In *International Conference on Machine Learning*, 9040–9051. PMLR.

Roscher, R.; Rußwurm, M.; Gevaert, C.; Kampffmeyer, M.; Dos Santos, J. A.; Vakalopoulou, M.; Hänsch, R.; Hansen, S.; Nogueira, K.; Prexl, J.; et al. 2024. Better, Not Just More: Data-Centric Machine Learning for Earth Observation. *IEEE Geoscience and Remote Sensing Magazine*.

Schick, T.; and Schütze, H. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352. Online: Association for Computational Linguistics.

Shi, J.; Ma, Q.; Ma, H.; and Li, L. 2024. Scaling Law for Time Series Forecasting. *arXiv preprint arXiv:2405.15124*.

SimpleMaps. 2024. SimpleMaps World Cities Database. Accessed April 16, 2024.

Source, M. O.; McFarland, M.; Emanuele, R.; Morris, D.; and Augspurger, T. 2022. microsoft/PlanetaryComputer: October 2022.

Sumbul, G.; Charfuelan, M.; Demir, B.; and Markl, V. 2019. BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 5901–5904. IEEE.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Tseng, G.; Cartuyvels, R.; Zvonkov, I.; Purohit, M.; Rolnick, D.; and Kerner, H. 2023. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. *arXiv preprint arXiv:2304.14065*.

Tseng, G.; Zvonkov, I.; Nakalembe, C. L.; and Kerner, H. 2021. Cropharvest: A Global Dataset for Crop-type Classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Wang, Y.; Braham, N. A. A.; Xiong, Z.; Liu, C.; Albrecht, C. M.; and Zhu, X. X. 2023. SSL4EO-S12: A Large-scale Multimodal, Multitemporal Dataset for Self-supervised Learning in Earth Observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3): 98–106.

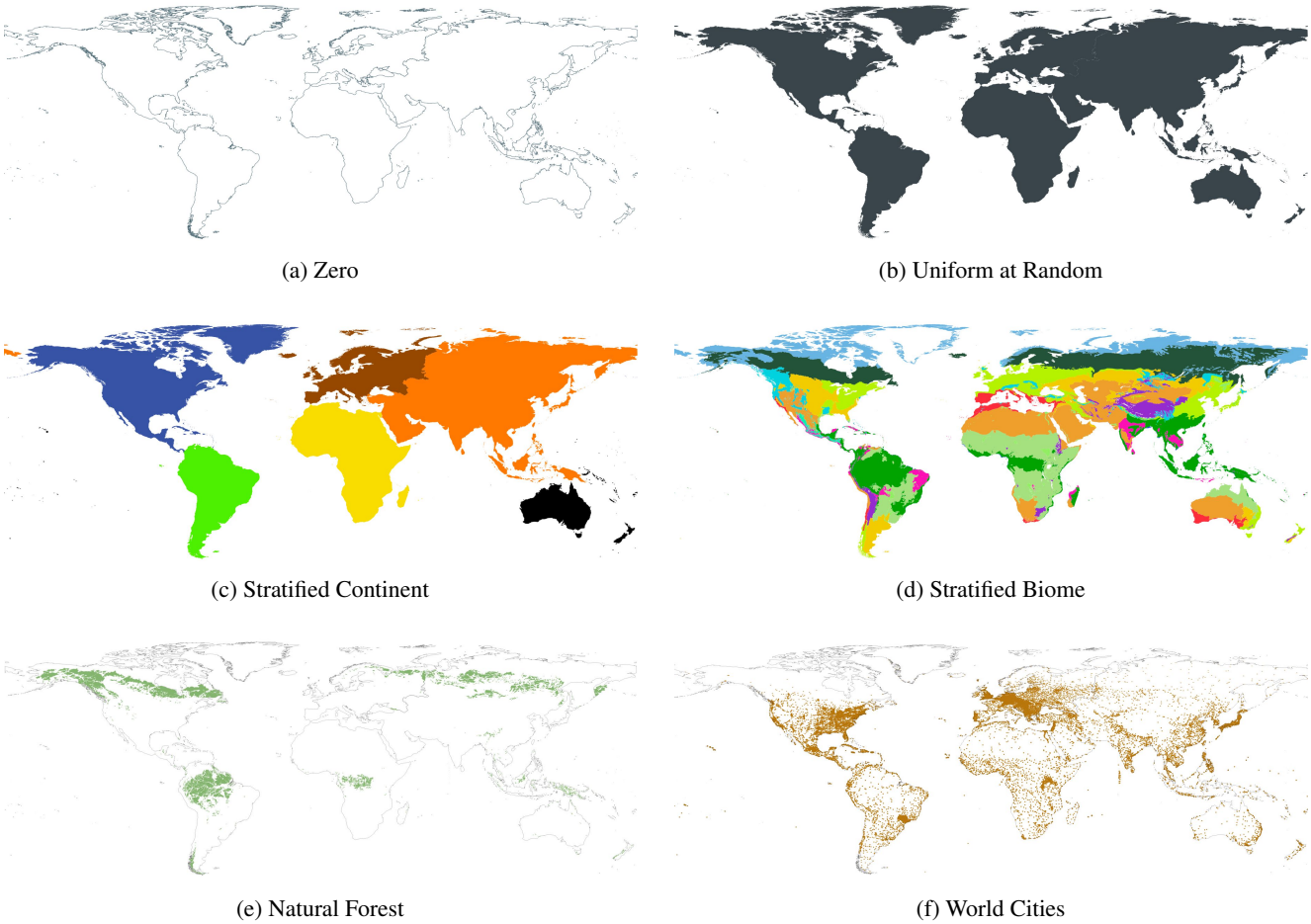
A Pre-training Data Distribution Study in General CV

The effect of pre-training data distribution on performance has been widely explored in general CV research. Hammoud et. al. shows that increasing pre-training data diversity improves the performance of the self-supervised model, but only when the distribution closely aligns with the downstream task data (Al Kader Hammoud et al. 2024). For an ImageNet-scale dataset, reducing the pre-training data by half (from 1M to 500k images) only decreases downstream performance by about 1-2% (Cole et al. 2022). Another study (Ramanujan et al. 2024) shows that pre-training with additional data enhances robustness, however, a significant gain can be achieved without large datasets. They show that using only 25K images from ImageNet or iNaturalist—6x smaller than the finetuning dataset—already provides noticeable improvements. Cole et. al. and Ramanujan et. al. establish that simply increasing data volume without increasing diversity does not always benefit. Longpre et al. (2024) examines how pre-training data choices affect language model performance, showing that data age, quality, toxicity filtering, and domain diversity significantly influence outcomes. Longpre et. al. also demonstrates that diverse data sources, like books and web content, improve robustness in language model development. Previous work also shows that different pre-training distributions lead to variations in transfer accuracy and these differences are more significant in few-shot finetuning (Entezari et al. 2023). Having a variety of objects and their characteristics in pre-training data significantly improves the generalization capability of the CLIP model (Abbasi, Rohban, and Baghshah 2024). Also, Fang et al. (2022) shows that CLIP's robustness largely depends on the choice of training data, with other factors having minimal impact.

B Details of Pre-training Data

As discussed in §3.3, to vary spatial distributions of pre-training datasets, we experimented with five data composition techniques each designed to capture varying levels of geographic coverage and regional characteristics as shown in Figure 3. These compositions range from broad global diversity to targeted environmental and urban-focused regions, offering unique spatial distributions for analysis. To do a fair comparison across pre-training sampling strategies, we used an equal number of pre-training samples in all data compositions for a given GFM. Details of pre-training data distribution for each GFM are as follows:

- **Presto**: For each composition, we resampled pre-training datasets from an existing pool of approximately $\sim$ 22 million samples provided by the Presto dataset (Tseng et al. 2023). Particularly, for each data composition, we sampled 6.5 million data points. The choice of 6.5 million samples was driven by the need to maintain equal samples across strata in stratified sampling techniques.
- **SatCLIP**: We created pre-training datasets by sampling 100,000 patches for each composition strategy, following the original SatCLIP data sampling configuration (Klemmer et al. 2023). We used Sentinel-2 data by excluding



(a) Zero

(b) Uniform at Random

(c) Stratified Continent

(d) Stratified Biome

(e) Natural Forest

(f) World Cities

Figure 3: Geographic distribution of pre-training data compositions

one band as specified in the SatCLIP implementation. To export data, we used Microsoft Planetary Computer catalog (Source et al. 2022). The temporal range for sampling spanned from January 1st, 2021 to September 15th, 2024. The dataset for each sampling strategy was created from scratch to ensure the uniqueness of each composition.

C Experiment Configuration

Pre-training: To pre-train Presto and SatCLIP, we used the same configuration and hyperparameters as specified in the original studies, utilizing the official repositories provided by Tseng et al. (2023)[1] and Klemmer et al. (2023)[2], respectively.

Downstream Tasks and Finetuning: As described in §3.2, we finetune the CropHarvest and EcoRegions tasks continent-wise. To achieve this, we first categorize the data samples of each task into continent categories, Table 2 provides the number of samples for all six continents and the total dataset size across both tasks. We then split the data from each continent into training, validation, and testing sets for finetuning and evaluation. We did not consider country-level finetuning and evaluation, as the country-level samples were insufficient, and the reduced sample size in the testing set prevented us from drawing reliable conclusions.

For CropHarvest, we modified the task from what was originally proposed in (Tseng et al. 2021). The original CropHarvest released an evaluation set focused on specific crop categories in three countries. Instead, we used the full CropHarvest dataset to create continent-specific subsets for finetuning. For each continent, we split the data into 70% for training, 10% for validation, and 20% for testing. For EcoRegions, we used the dataset including its training, validation, and testing splits as it was employed in SatCLIP.

As discussed in the §4, we conducted finetuning and evaluation on both tasks using an increasing amount of data. For both downstream tasks, we finetuned the model on varying sample sizes, starting from $n = 10$ up to the total available training samples in the continent. As the number of fineuning samples n increases, we proportionally reduce the number of iterations r used to compute the final average performance metric, i.e., $n \propto \frac{1}{r}$. Importantly, in all finetuning and evaluation, only the training set is changing, validation and test set remain fixed across all experimental configurations.

[1]https://github.com/nasaharvest/presto

[2]https://github.com/microsoft/satclip

Table 2: Total and continent-wise number of samples in downstream tasks

	AF	**AS**	**EU**	**NA**	**OC**	**SA**	**Total**
CropHarvest	21693	14475	15784	10656	756	6812	70176
EcoRegions	4027	8333	2246	5747	1177	2344	23874

D Additional Results

As mentioned in §3.1, we evaluated each task using both parametric models (Multi-Layer Perceptron (MLP) and logistic regression) and non-parametric models (Random Forest (RF) and K-Nearest Neighbors (KNN)).

Figure 4, 5, and 6 present line plots of the F1 score with confidence intervals for CropHarvest, finetuned on RF, KNN, and Regression, respectively. Similarly, results for EcoRegions finetuned using RF, KNN, and MLP are shown in Figure 7, 8, and 9, respectively. Across all these line plots, we observe that performance improves for all continents as the number of samples increases. Moreover, findings reported in the main paper remain consistent: the impact of differences in data composition diminishes as the sample size grows.

E Limitations and Future Work

This study focuses primarily on providing preliminary insights into the impact of pre-training data distribution on the performance of GFMs. There is significant room for extending this research that helps in development of robust GFMs. Below, we outline key areas for future exploration:

Our analysis includes two GFMs and a single downstream task per GFM. Future work can be expanded to evaluate a broader range of GFMs and multiple downstream tasks. This will help to strengthen the findings reported in this research and provide guidelines for pre-training data sampling techniques tailored to diverse EO applications.

In terms of describing geographic diversity, this study selected groupings based on continents, biomes, cities, and natural forests. However, there are many other factors such as countries, hemispheres, climate zones, land cover/land use, etc can also be used which raises an important question: *What groupings should we define pre-training data diversity with respect to?*

Our experiments are conducted at the continent-level due to the infeasibility of country-level finetuning and evaluation (details in Appendix C). Future work can consider downstream tasks with well-distributed data across countries, enabling country-level evaluations and exploring the pre-training data distributions best suited for such granular analyses.

Although this work focuses on balanced and stratified sampling methods for continent and biome grouping, future studies can explore scenarios where intentionally (it can be somewhat or totally) biased sampling techniques can enhance the performance for specific regional tasks. This will help to understand the trade-offs between global coverage and regional specialization in GFM development.

While this paper focuses on the impact of geographic diversity of pre-training data, the influence of data quantity is another crucial factor. Future research can investigate the relationship between the volume of pre-training data and model performance, exploring whether increasing the amount of data leads to proportional improvements in downstream task performance.

By addressing these gaps, the GeoAI community can develop more effective GFMs and establish best practices for pre-training data selection, ultimately advancing machine learning solutions for Earth observation tasks.

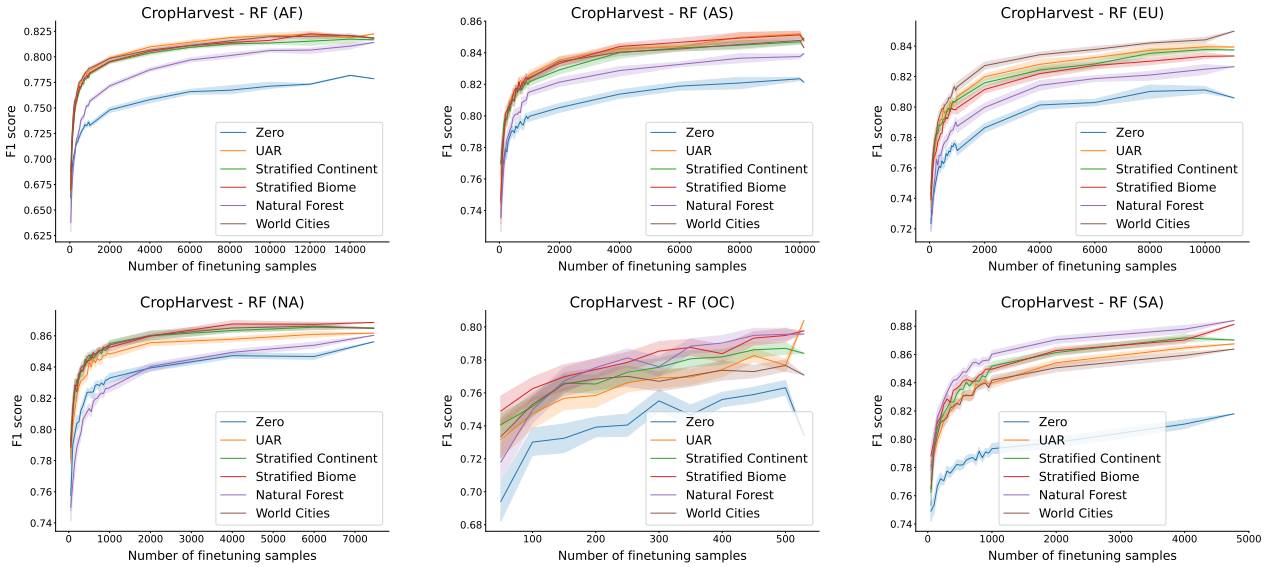


Figure 4: Continent-wise results for CropHarvest task with Random Forest

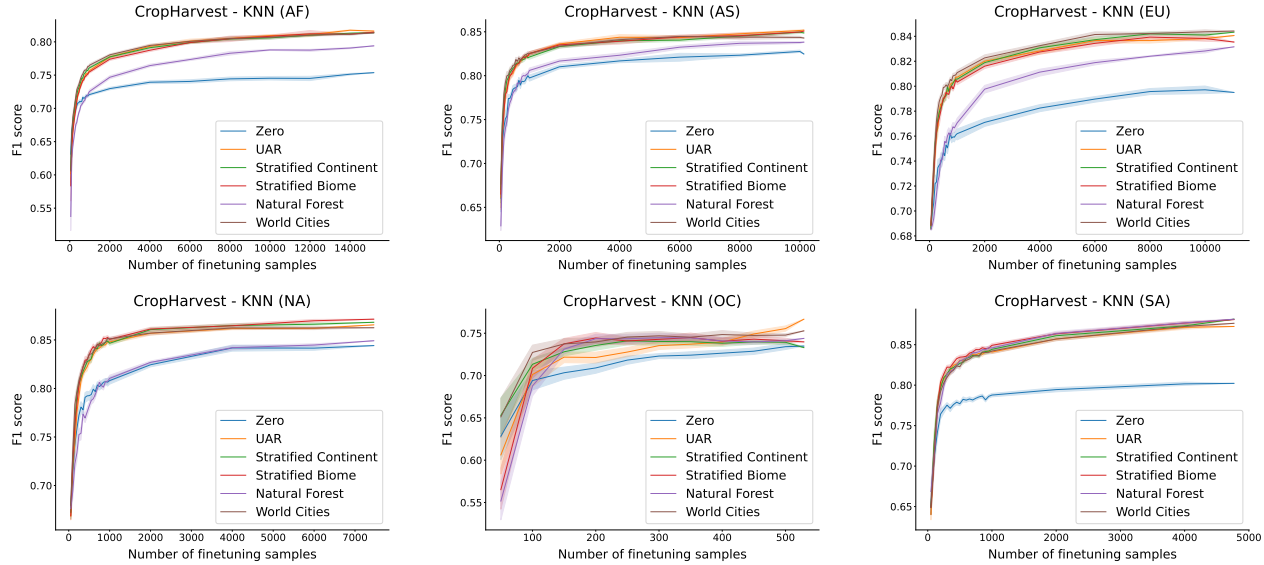


Figure 5: Continent-wise results for CropHarvest task with KNN

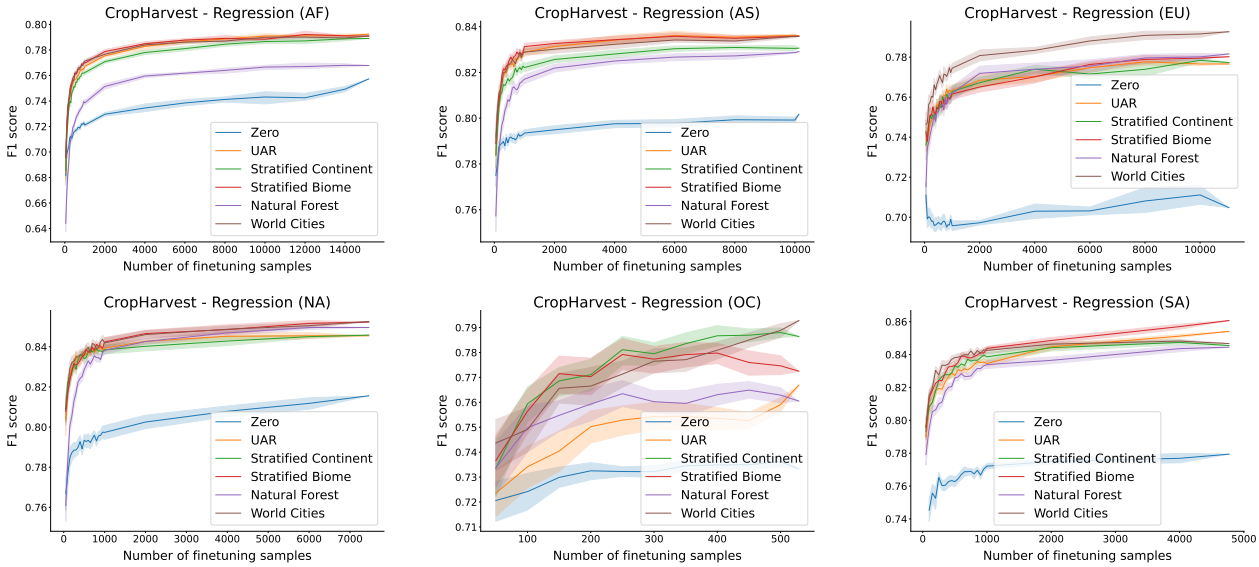


Figure 6: Continent-wise results for CropHarvest task with Regression

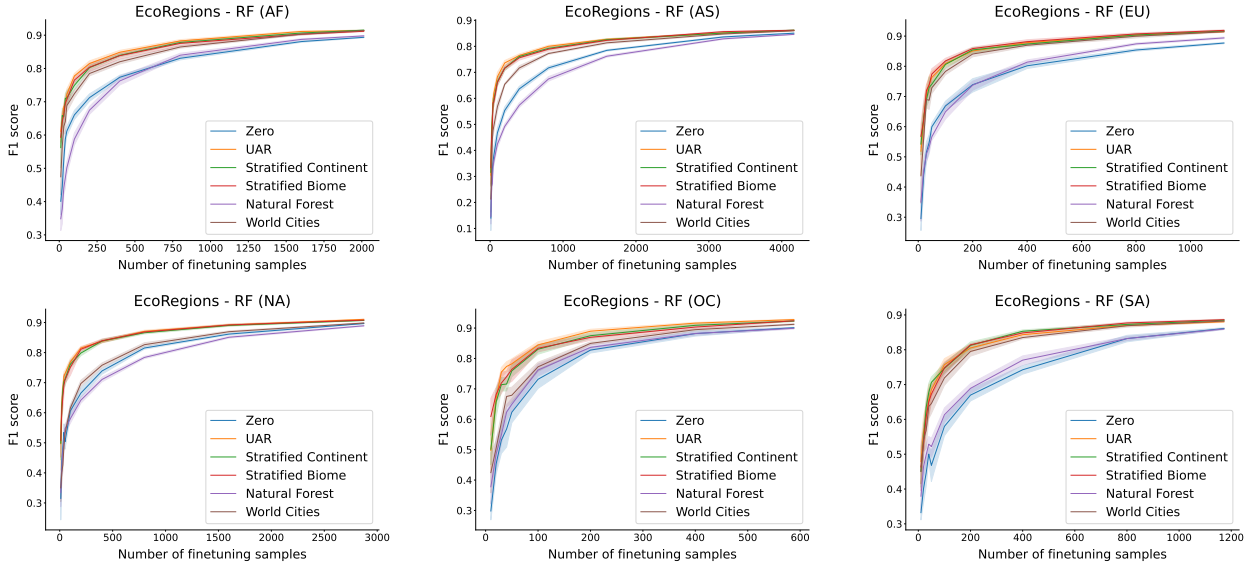


Figure 7: Continent-wise results for EcoRegions task with Random Forest

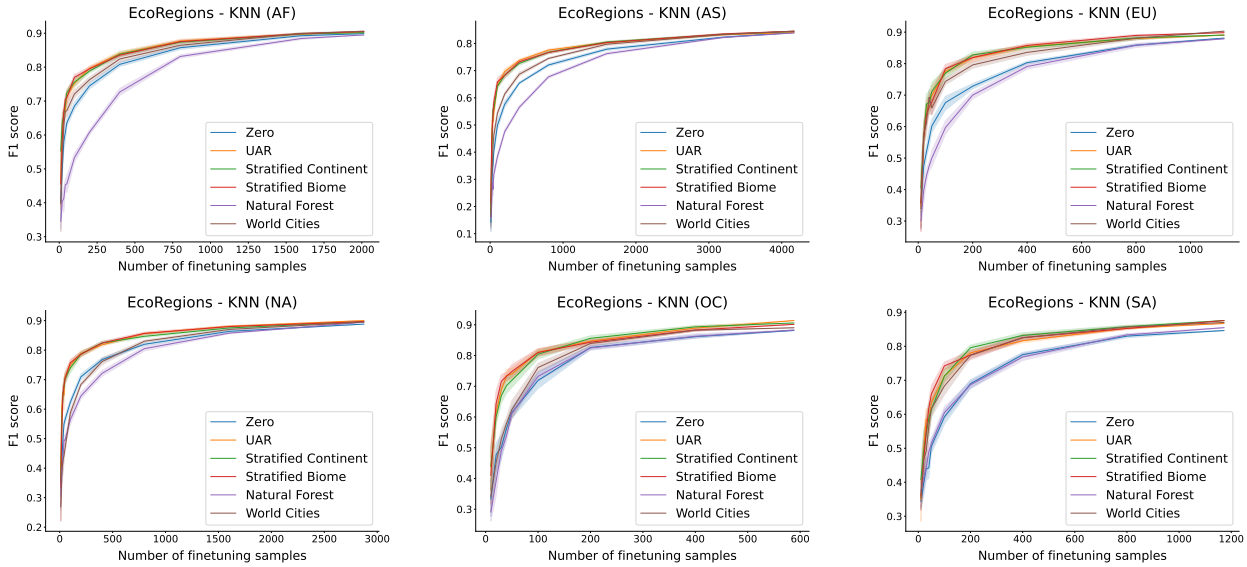


Figure 8: Continent-wise results for EcoRegions task with KNN

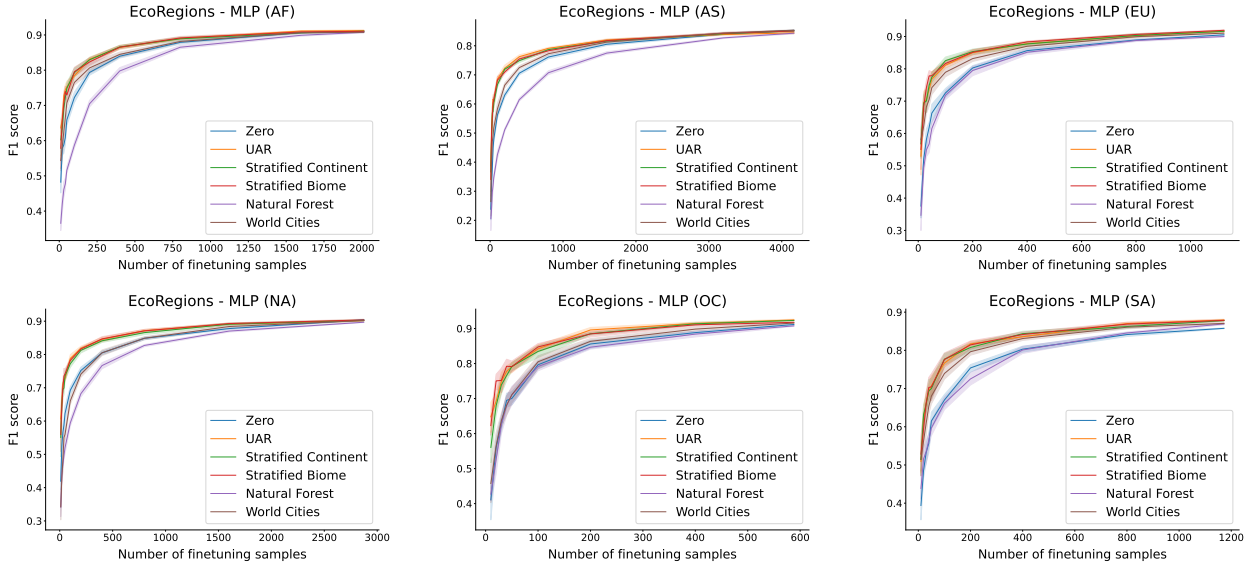


Figure 9: Continent-wise results for EcoRegions task with MLP