Uncertainty and Generalizability in Foundation Models for Earth Observation

Raúl Ramos-Pollán
Universidad de Antioquia, Colombia
raul.ramos@udea.edu.co

Freddie Kalaitzis
University of Oxford, UK
freddie.kalaitzis@cs.ox.ac.uk

Karthick Panner Selvam
University of Luxembourg
karthick.pannerselvam@uni.lu

Abstract

We take the perspective in which we want to design a downstream task (such as estimating vegetation coverage) on a certain area of interest (AOI) with a limited labeling budget. By leveraging an existing Foundation Model (FM) we must decide whether we train a downstream model on a different but label-rich AOI hoping it generalizes to our AOI, or we split labels in our AOI for training and validating. In either case, we face choices concerning what FM to use, how to sample our AOI for labeling, etc. which affect both the performance and uncertainty of the results. In this work, we perform a large ablative study using eight existing FMs on either Sentinel 1 or Sentinel 2 as input data, and the classes from the ESA World Cover product as downstream tasks across eleven AOIs. We do repeated sampling and training, resulting in an ablation of some 500K simple linear regression models. Our results show both the limits of spatial generalizability across AOIs and the power of FMs where we are able to get over 0.9 correlation coefficient between predictions and targets on different chip level predictive tasks. And still, performance and uncertainty vary greatly across AOIs, tasks and FMs. We believe this is a key issue in practice, because there are many design decisions behind each FM and downstream task (input modalities, sampling, architectures, pretraining, etc.) and usually a downstream task designer is aware of and can decide upon a few of them. Through this work, we advocate for the usage of the methodology herein described (large ablations on reference global labels and simple probes), both when publishing new FMs, and to make informed decisions when designing downstream tasks to use them.

1 Introduction

Foundation Models (FMs) have been proven particularly useful in scarce label scenarios, where models are built with a limited amount of labeled data for different downstream tasks. Earth Observation (EO) is not an exception, and EO labeled data comes with its own particularities. It is usually quite heterogeneous, being available in the particular regions where the labelling effort is focused for each specific downstream task. Furthermore, they are more abundant in regions like the US and Europe where science resources are larger.

In this work, we assume we have a limited labeling budget on the AOI for the downstream task of interest, the **target AOI** and we want to make the best use of it by using an FM. We consider two cases: (1) there is an **external AOI** having a healthy amount of labels to train a downstream model

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

for our task, therefore we would only need to use our label budget in the target AOI for validation; and (2) we split our labeling effort in our target AOI both for training a model and validating it.

We are firstly concerned with **spatial generalizability**, dealing with models trained in one AOI doing inference somewhere else. Performance can be compromised due to many reasons, because land features are simply different across AOIs (vegetation, buildings, etc.), because the FM did not capture enough information, because we are not training with sufficient data, etc. Secondly we are concerned with the **uncertainty** stemming from small data scenarios (epistemic uncertainty, see [1]).

We test eight different FMs using Sentinel 1 (S1, Synthetic Aperture Radar -SAR-) amplitude and Sentinel 2 (S2, optical multi-spectral) as input data to obtain chip level embeddings. Then, we use the ESA World Cover product (`esawc`) as global labels for a chip level percentage estimation downstream task using linear regression for each dataset class (tree cover, built up, permanent water, etc.) so that we can perform ablations and tests across the globe. With this we hope to provide a first indication on what to expect when applying FMs to related downstream tasks (for instance biomass estimation, human footprint, etc.)

We use a simple linear regression (LR) model to make a large ablation study (over 500K models trained) using the US, Europe and China as external AOIs and Colombia, Perú, India, Kenya, California, Texas, Spain and Germany as target AOIs. LR provides us both computational affordability for such a large study and a glimpse on the straight forward information content provided by FM embeddings.

The results below show a wide diversity of situations. Generalizability is higher in some downstream tasks and AOIs than in others; as we increase the labeling effort in target AOIs we reduce the uncertainty of our experiments on different degrees for different FMs and tasks; different sampling methods provide different outcomes, etc. But in general we see that all FMs behave similarly, with certain advantages of one FM over the other for certain specific situations. And yet, it is surprising the level of predictive power we can obtain in several cases just using a linear probe directly on FM embeddings.

In all, there are many design decisions behind each FM and downstream task (input modalities, data sampling, architectures, pretraining method, etc.) and a downstream task designer is usually aware and can decide upon a few of them. Through this work, we advocate for the usage of the methodology herein described (reference global labels and simple probes), both when publishing new FMs, and to make informed decisions when designing downstream tasks to use them.

2 Previous works

During the last few years there has been a frantic effort to develop a wide range of Earth Observation Foundation Models, covering different input sensors, formats, time series, world regions, etc. See for instance [2] [3] [4] [5] [6] [7] [8] [9] just to name a few. Vision Transformers (ViT) [10] as architecture and masked autoencoders based losses [11] dominate this landscape, and most FMs draw insights and methods from the self supervised community [12].

Several reviews pinpoint many aspects and challenges for FMs in Earth Observation [13] [14]. Key aspects among those challenges are **engineering** (how easy it is to inject my data into FMs), **semantics** (how sensible are FMs to the earth features I am interested in), **multimodality** (how well can the FM exploit inputs from multiple sensors), **flexibility** (how robust are FMs to missing data and different time and spatial resolutions), **generalizability** (can I use FMs to build models doing inference across regions?) and **uncertainty** (how much variance should I expect when feeding small sampled data to FMs?). In this work we address the last two.

Finally, some efforts have focused on creating benchmarks or benchmarking datasets for EO. See for instance [15] [16]. The focus of this work is complementary, since we are trying to obtain a comprehensive view of FMs generalizability and uncertainty rather than focusing on complex end-to-end downstream tasks.

AOI	usage	Mkm^2	number of chips	landmass coverage
US	external	7.9	30418	10%
Europe	external	5.5	35936	17%
China	external	9.4	35934	10%
Colombia	target	1.2	4222	10%
Peru	target	1.3	4860	10%
India	target	3.1	11875	10%
Kenya	target	0.6	2228	10%
Spain	target	0.5	1905	10%
Germany	target	0.5	6556	48%
California	target	0.4	1589	10%
Texas	target	0.7	2620	10%

Table 1: Areas of Interest (AOIs) used in this work. **External** AOIs are only used for training downstream models. **Target** AOIs are used both for training and validation. We sampled randomly 10% of the world landmass, except in Europe where we sampled 100% on an area of 1000km centered in Luxembourg.

3 Foundation models, data and downstream tasks

Areas of Interest (AOIs) Table 1 details the AOIs used in this work. We selected three **external AOIs** (US, Europe, China) typically rich on labels upon which we train models; and test them in the **target AOIs** (Colombia, Peru, Kenya, India, Spain, Germany, California, Texas). They were selected to intuitively represent a diversity of terrain features, which can be observed in Figure 1. The inclusion of European countries and US states in the target AOIs is to have a sense on how region wide models perform on included regions.

Input data to FMs We use chips of size 512×512 pixels which, at Sentinels resolution of 10m/pixels, corresponds to a spatial resolution of 5.12km$\times$5.12km, or 26.2km^2. We created S1 and S2 datasets using `geetiles`[1] which pulls data from Google Earth Engine [17]. We took data for a full year and computed the median per pixel per season (winter, spring, summer, fall). For S1 this results in 16 channels (vv/vh[2], ascending/descending per season), and for S2 this results in 44 channels (11 bands per seasson). The S2 bands used are [B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12] as accepted by all S2 models.

Foundation Models We used eight foundation models as detailed in Table 2. They were selected using criteria of easiness to use, similar model size and pretrained on Sentinel imagery. Each FM either accepts S1 or S2 chips as input data and produces an embeddings vector for each chip. This embeddings vector is then used as input to the linear models detailed below. So, for each of our AOIs there a set of embeddings vectors for all its chips for each FM. Note that different FMs are trained on different world regions and/or sampling methods.

Downstream tasks We use the ESA World Cover product [18] (`esawc`) which delivers 11 landcover classes at the same spatial resolution as S1 and S2 (10m/pixel). Out of those we focus only on the seven classes in Figure 1, since the rest occur very rarely, and the figure shows their distribution across the selected AOIs. For each of those classes, we set up a chip-level regression task to predict the percentage of pixels in the chip labeled with that class. This is a number between 0 and 1 per chip.

4 Methods

Linear regression on chip level percentage We train a linear regression model for each experimental setup described below (`esawc` class, train AOI, target AOI, etc.) By just using linear regression we intend to have a sense on the *raw* information content provided by FMs' embeddings with respect to the different `esawc` classes. Also, with the large number of experiment combinations due to the ab-

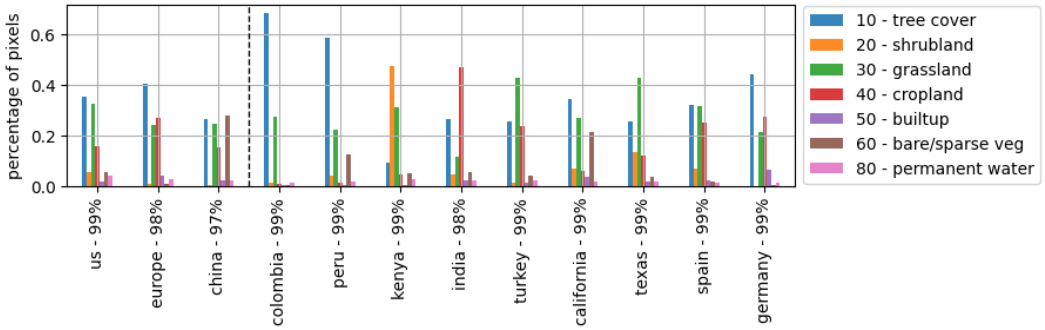
[1] `http://github.com/rramosp/geetiles`

[2] SAR emitting a vertically polarized signal and receiving both vertical (vv) and horizontal (vh)

Foundation Model	emb. size	params	arch.	pretrain method	AOI used to train the FM
Sentinel 1 models					
`s1-fdl2024mae` [7] ⋆	768	95M	ViT	MAE	world 10% random
`s1-fdl2024clip` [8] ⋆	768	90M	ViT	CLIP	world 10% random
`s1-clay_v02` †♮	768	113M	ViT	MAE	world landcover stratified
`s1-clay_v1` †♮	768	201M	ViT	MAE	world landcover stratified
Sentinel 2 models					
`s2-fdl2024mae` [7] ⋆	768	98M	ViT	MAE	world 10% random
`s2-fdl2024clip` [7] ⋆	768	91M	ViT	CLIP	world 10% random
`s2-clay_v02` †♮	768	113M	ViT	MAE	world landcover stratified
`s2-prithvi` [2] ‡	768	116M	ViT	MAE	US climate stratified

Table 2: Foundation Models (FMs) used in this work. FMs with ⋆ take as input separated median per pixel per season (8 channels for vv, vh for seasons for S1, 11 channels for S2) to produce a single embeddings vector. FMs with † are fed sequentially each season chip, producing four embeddings vectors which are then averaged into a single one. Both S1 and S2 CLIP encoders were trained together. Prithvi (‡) accepts three time steps so we input the first three season medians (winter, spring, summer). `clay_v1` also accepts S2 data but was left out because of timing constraints. Clay models (♮) can be found at `https://github.com/Clay-foundation/model`

Figure 1: Distributions of the seven selected `esawc` classes for this work on each AOI. Percentage number indicate how much of the AOI landmass is covered by those seven classes. Dotted line separates external AOIs from target AOIs. Since we use 7 out of the 11 classes originally present in `esawc`, the percentages beside each country name represent what portion of the country is covered by those 7 classes.

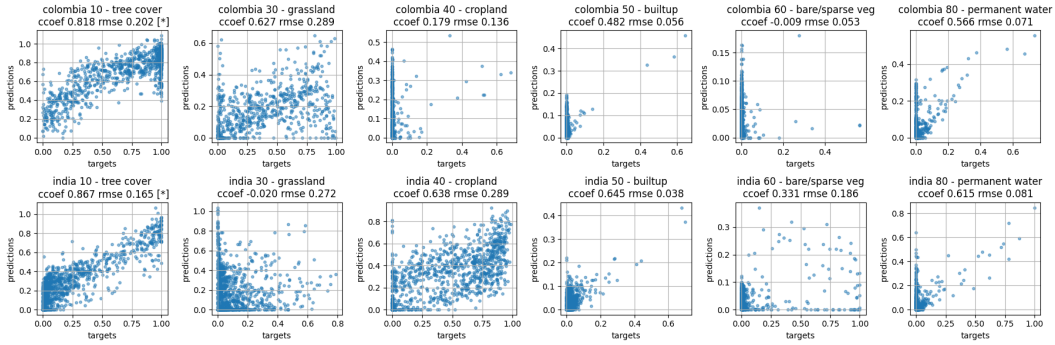


lations shown below, linear regression seems a reasonable choice to make this work computationally feasible.

Metric We use the correlation coefficient between predictions and targets. Observe how, in Figure 2 the correlation coefficient seems to capture better than RMSE whether the linear probe on the embeddings is actually able to perform the task. Notice, for instance, Colombia on class `permanent water` with misleading low RMSE, or India on `tree cover` vs. `bare/sparse vegetation` with similar RMSE, but the model clearly picking up the first class, but not the second. For interpretation of the results we establish a threshold of 0.7 above which we consider FM embeddings are actually picking up meaningful information on the underlying predictive task.

Sampling We use four different sampling methods when selecting data for training, both when training with the external AOIs and the target AOIs. This sampling is done on the datasets described in Table 1. With **esawc sampling** we sample proportional to distribution of `esawc` classes in the AOI, so that the presence of all classes is as uniform as possible. With **fps sampling** we use Furthest Point Sampling [19] on the embeddings for each FM using euclidean distance, which intuitively favors an overall variance of the embeddings. With **random sampling** we do a spatially uniform random

Figure 2: Example predictions of a linear probe on the embedding from FM `s1-fdl2024-mae`, trained with data from Europe on different `esawc` tasks and target AOIs. An asterisk `[*]` denotes a correlation coefficient greater than 0.7 as the threshold above which we will consider the embeddings do contain useful information for that task and target AOI.



sampling. And with **srtm sapling** we do sampling proportional to the mean elevation of each chip so that we get as many different elevations as possible.

Experimentation We did two sets of ablations (1) using external AOIs for training models, and testing on target AOIs; and (2) using part of the data in the target AOIs for training, and testing on the remaining target AOI data. We ablated on the following hyperparameters: FM embeddings (see Table 2), external and target AOIs (see Table 1), number of train elements on external AOI [300,3000,30000], number of train elements on the target AOI [10,50,100,500], number of test elements on the target AOI [10,50,100,500] and `esawc` class (7 classes).

This results in $\sim$18K models when using external AOIs for training, and $\sim$7K models when splitting target AOIs for train and test. In this later case, we don't do cross AOI training and testing. We repeat each experiment 20 times, resampling train and test data each time and reporting the mean and standard deviation of the metric. Sampling test data is always done using spatially uniform random sampling. With this, in total we trained some 500K linear regression models.

5 Results

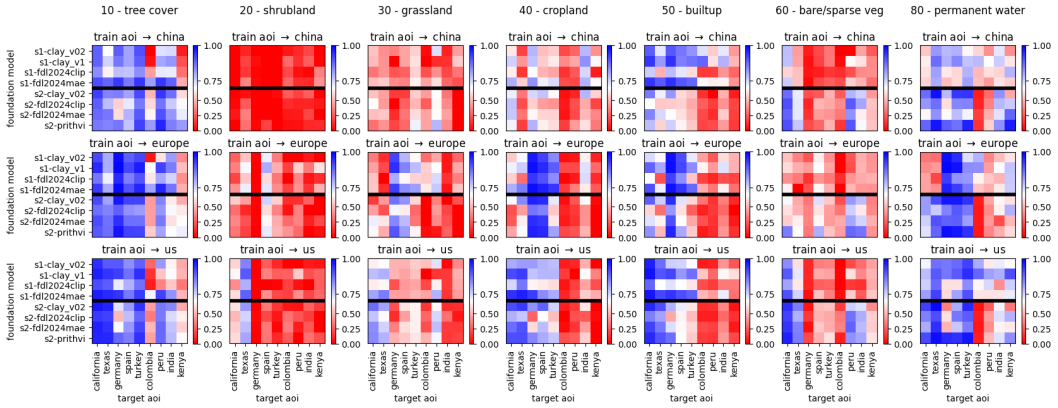
Foundation Models on different target AOIs and tasks Figure 3 shows an overall view of FM embeddings performance when trained on the different external AOIs and tested on our target AOIs. We can already observe several aspects. The US and Europe have reasonable generalizability across their regions in most tasks (observe, the first four columns on each chart corresponding to US on Spain and Germany, and Europe on California and Texas)

Tasks `tree cover` and `permanent water` show reasonable generalizability to non US and non European AOIs. Models for `permanent water` trained in Europe seem to work better in California and Texas only if we use Sentinel 2.

Some tasks seem very AOI specific. In task `cropland` we get some signal when transferring models between the US, Europe and Turkey. And also in models trained in China applied in India. We believe this might be due to the inherent differences between AOIs (different crops around the world) and these classes probably gather a large variety of land features (like crops of palm trees look very different from crops of soy). Task `shrubland` shows very low performance and no spatial generalizability at all and task `grassland` only timidly within the US and within Europe.

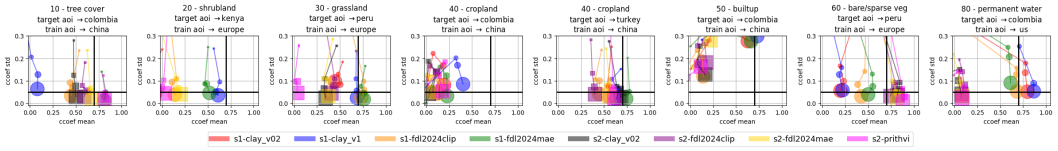
Other tasks seem inherently dependent on the sensor. For instance, generalizability in Perú and India for `bare/sparse vegetation` seems to occur when using Sentinel 2 FMs regardless where the models were trained, and `builtup` seems more generalizable with Sentinel 1 (bluer above the black horizontal line), which is in line with the expected interaction of a SAR signal with man made objects.

Figure 3: Overall view of linear probes with different train AOIs, target AOIs and dowstream tasks, showing models trained on 30K elements in external AOIs and tested with 500 elements in target AOIs. Showing the mean of 20 runs. Correlation threshold is set at 0.7 (white). Bluer positions represent greater correlation between predictions and targets, redder ones worse. Black horizontal line splits S1 and S2 FMs.



Uncertainty in target AOIs when training with external data Figure 4 shows some of the ablations on the number of elements used in target AOIs to test models trained on the external AOIs. Each chart corresponds to a column in Figure 3 and they were selected as they seem border cases seldom overcoming the 0.7 correlation coefficient threshold we established. Since we assume we are on a limited labeling budget on target AOIs, we are interested in having the minimum amount of labels without loosing overall predictive performance and, most importantly, with metric stability. Since we are repeating the same data and model configuration 20 times, resampling every time, we measure both the mean performance (correlation coefficient) and its standard deviation. Therefore, we also establish a threshold on the correlation coefficient standard deviation at 0.05.

Figure 4: Selected ablations increasing the number of elements for test chips in the target AOI used represented with **dot size** in the set of values [10,50,100,500]. **Squared markers** represent models with Sentinel-2 input, round ones with Sentinel-1 input. Thresholds of 0.7 correlation coefficient mean and 0.05 standard deviation are shown

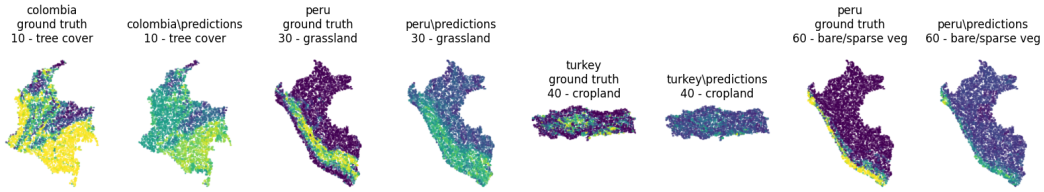


Then, Table 3 details the cases we consider would be useful in practice with these thresholds (0.7 mean, 0.05 stdev). Observe that for some cases we are not able to get satisfactoyy FMs (although class `permanent water` for Colombia is left barely out with an standard deviation of 0.052). And for `builtup` in Colombia, even if we overcome the correlation coefficient threshold for its mean with 500 chips labeled, the uncertainty is quite high (around 0.3). Figure 5 shows those predictions.

Finally, observe how the modality dependancy is much more explicit in several cases in Figure 4 having well differentiated the performance of FMs using as input data S1 (circles) and S2 (squares).

Uncertainty in target AOIs when training with its own data We now consider whether labeling budget on the target AOI can be split into train and validation and it is worthwhile as compared to using external AOIs data for training as shown above. Figure 6 shows the correlation coefficient mean and standard deviation obtained as we ablate on the train and test split on each one of the selected target AOIs. We discriminate according to the sampling method used when training.

Figure 5: Predictions for one run of cases in Table 3. Color shows the percentage of the `esawc` class, either the target or the prediction.



target aoi	external aoi	esawc class	FM	test elems	corr coef mean	corr coef std
colombia	china	tree cover	s2-prithvi	50	0.814	0.048
peru	europe	grassland	s1-fdl2024mae	100	0.764	0.045
turkey	china	cropland	s1-fdl2024mae	100	0.740	0.046
peru	europe	bare/sparse veg	s2-prithvi	100	0.894	0.030

Table 3: Cases in the bottom right quadrants in Figure 4 with the correlation coefficient mean is greater than 0.7, its standard deviation less than 0.05 and the number of test elements is lowest.

From those results, Table 4 shows the details of the cases where the correlation coefficient mean and standard deviation lies within the established thresholds. Observe first that, as could be expected, there is an overall increase in performance in two cases (`shrubland` in Kenya and `permanent water` in Colombia) and now we are within the performance thresholds. Observe that, in all cases, we can obtain high correlation coefficients at reduced uncertainty (standard deviation) with a handful of labeled chips split between train and test.

train aoi	esawc class	FM	number elems total (test/train)	sampling	corr coef
colombia†	tree cover	s1-fdl2024mae	110 (10/100)	fps	0.947 $\pm$0.032
colombia‡	tree cover	s1-fdl2024mae	60 (50/10)	random	0.816 $\pm$0.048
kenya†	shrubland	s1-clay_v1	100 (50/50)	random	0.923 $\pm$0.023
kenya‡	shrubland	s1-clay_v1	60 (50/10)	fps	0.791 $\pm$0.049
peru†	grassland	s1-fdl2024mae	100 (50/50)	random	0.888 $\pm$0.027
peru‡	grassland	s1-fdl2024mae	60 (50/10)	esawc	0.796 $\pm$0.050
turkey†	cropland	s1-fdl2024mae	110 (10/100)	fps	0.939 $\pm$0.035
turkey‡	cropland	s1-fdl2024mae	60 (50/10)	fps	0.836 $\pm$0.038
peru†	bare/sparse veg	s2-prithvi	150 (50/100)	fps	0.949 $\pm$0.032
peru‡	bare/sparse veg	s2-prithvi	60 (50/10)	esawc	0.896 $\pm$0.028
colombia†	permanent water	s1-clay_v1	550 (500/50)	fps	0.909 $\pm$0.035
colombia‡	permanent water	s1-clay_v1	150 (100/50)	random	0.905 $\pm$0.047

Table 4: Cases in the bottom right quadrants for charts in Figure 6 where the correlation coefficient mean is greater than 0.7 and its standard deviation less than 0.05. Cases with $\ddagger$ are the ones with the least number of total elements, and with † are the ones where the correlation coefficient was greatest.

Sampling methods We also make a reflection on the different sampling methods we used to select which data to use for training (whether on external or target AOIs). Observe how in Table 4 the sampling method becomes relevant when selecting the best performing cases. In particular Furthest Point Sampling (FPS) becomes a valuable alternative in the majority of the cases. Recall that FPS samples differently on each embeddings space, attempting to promote large euclidean distances within the embeddings of the sampled chips. These sampling differences also become particularly important when using external AOIs for training. Observe how in Figure 7 with fewer elements (300 or 3000), the sampling method has great influence.

Figure 6: Same ablations as in Figure 4 but using the same target AOI to split the data between train and validation, and using only the FM with which the best performance was obtained in Figure 4. **Dot size** represents the number of labeled chips for validation in the set [10,50,100,500] from smaller to larger. **Shading** represents the number of labeled chips used for train within the target AOI in the set [10,50,100,500] from lighter to darker.

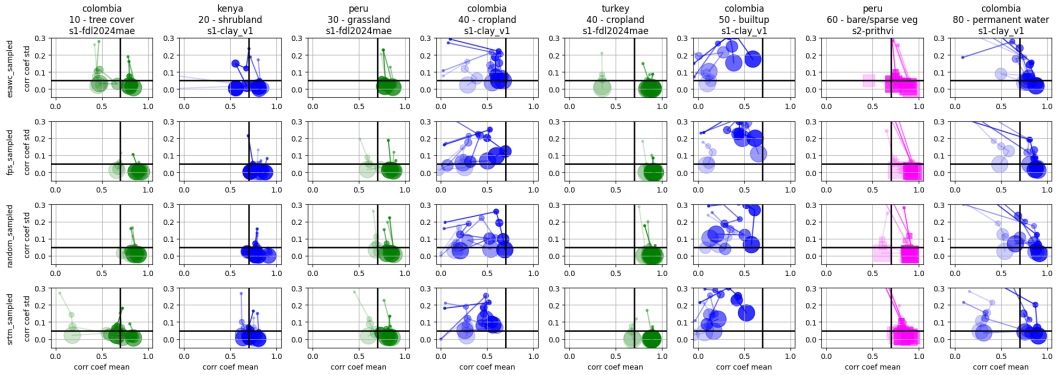
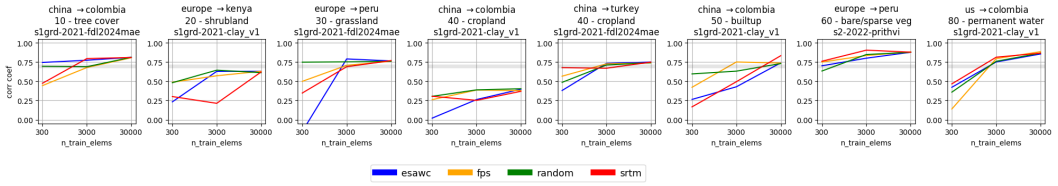


Figure 7: Effect of different sampling methods when we ablate on the number of train elements when using external AOIs for training models. Showing the mean of 20 repetitions for each configuration.



6 Conclusion

Through this study we have seen how generalizability and uncertainty vary greatly across different combinations of AOIs, tasks and input modalities. This shows that, when facing a labeling effort for a downstream task, it is key to make the right selection of FM and sampling method to make an efficient use of the labeling budget at hand.

As a result, we advocate for worldwide studies with **simple methods**, in which for each FM embeddings a number of **representative downstream tasks** globally available are tested against large combinatorics of train and target AOIs (countries, regions, continents, etc.), measuring generalizability and uncertainty with the methodology illustrated in this work. With proper experimentation such representative downstream tasks would be proxies of many other tasks related to them for which there are no global labels. Such studies would enable (1) better comparisons between FMs; and (2) pinpoint decisions on how to design downstream tasks within labeling budget constraints.

Acknowledgments and Disclosure of Funding

This work has been enabled by FDL Europe | Earth Systems Lab (https://fdleurope.org) a public / private partnership between the European Space Agency (ESA), Luxembourg Space Agency, Trillium Technologies, the University of Oxford in partnership with Google Cloud, NVIDIA Corporation, RSS Hydro, LuxProvide. We are thankful to the SAR-FM FDL 2024 Team and all the reviewers that participated in it.

References

[1] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[2] Johannes Jakubik et al. Foundation models for generalist geospatial artificial intelligence, 2023. URL https://arxiv.org/abs/2310.18660.

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211, 2022.

[4] Adam J Stewart, Nils Lehmann, Isaac A Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *arXiv preprint arXiv:2306.09424*, 2023.

[5] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[6] Michael J. Smith, Luke Fleming, and James E. Geach. Earthpt: a time series foundation model for earth observation, 2024. URL https://arxiv.org/abs/2309.07207.

[7] Matt Allen, Francisco Dorr, Joseph A. Gallego-Mejia, Laura Martínez-Ferrer, Anna Jungbluth, Freddie Kalaitzis, and Raúl Ramos-Pollán. Large scale masked autoencoding for reducing label requirements on sar data, 2023. URL https://arxiv.org/abs/2310.00826.

[8] Matt Allen, Francisco Dorr, Joseph A. Gallego-Mejia, Laura Martínez-Ferrer, Anna Jungbluth, Freddie Kalaitzis, and Raúl Ramos-Pollán. Fewshot learning on global multimodal embeddings for earth observation tasks, 2023. URL https://arxiv.org/abs/2310.00119.

[9] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs].

[12] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022.

[13] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A Wernke, and Yuankai Huo. Ai foundation models in remote sensing: A survey, 2024. URL https://arxiv.org/abs/2408.03464.

[14] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[15] Casper Fibaek, Luke Camilleri, Andreas Luyts, Nikolaos Dionelis, and Bertrand Le Saux. Phileo bench: Evaluating geo-spatial foundation models, 2024. URL https://arxiv.org/abs/2401.04464.

[16] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: Toward foundation models for earth monitoring, 2023. URL https://arxiv.org/abs/2306.03831.

[17] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2017.06.031. URL https://www.sciencedirect.com/science/article/pii/S0034425717302900. Big Remotely Sensed Data: tools, applications and experiences.

[18] Daniele Zanaga et al. Esa worldcover 10 m 2020 v100, October 2021. URL https://doi.org/10.5281/zenodo.5571936.

[19] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017.