The View From Space: Navigating Instrumentation Differences with EOFMs

Ryan P. Demilt[*], **Nicholas LaHaye, Karis Tenneson**
Spatial Informatics Group, Pleasanton, CA 94566, USA

Abstract

Earth Observation Foundation Models (EOFMs) have exploded in prevalence as tools for processing the massive volumes of remotely sensed and other earth observation data, and for delivering impact on the many essential earth monitoring tasks. An emerging trend posits using the outputs of pre-trained models as 'embeddings' which summarize high dimensional data to be used for generic tasks such as similarity search and content-specific queries. However, most EOFM models are trained only on single modalities of data and then applied or benchmarked by matching bands across different modalities. It is not clear from existing work what impact diverse sensor architectures have on the internal representations of the present suite of EOFMs. We show in this work that the representation space of EOFMs is highly sensitive to sensor architecture and that understanding this difference gives a vital perspective on the pitfalls of current EOFM design and signals for how to move forward as model developers, users, and a community guided by robust remote-sensing science.

1 Introduction

The remote sensing community has long driven impact through the skillful adoption and adaptation of machine learning techniques Ma u. a. (2019), particularly owing to the close relationship of many satellite image processing problems to developments in the computer vision domain. Recently, a paradigm shift has been proposed where in distinction to bespoke architectures and training procedures for each new task, models are proposed that are pre-trained on large collections, often entire mission archives worth, of data using self-supervised objectives. The data collections and objectives are carefully selected with the goal of learning discriminative, high-level but general features, producing 'Foundation Models' which have generic utility and can be used for a variety of downstream tasks. A diversity of models, largely based on variants of the ViT Dosovitskiy u. a. (2021) have been proposed for the task Fuller u. a. (2023); Wang u. a. (2023); Jakubik u. a. (2023); Xiong u. a. (2024); Clay (2024); Jakubik u. a. (2025), each with bespoke data curation regimes, architectural modifications, and pretraining objectives; often with only one optical input stream used in pretraining.

In contrast to the relatively uniform space of traditional computer vision images, which share a consistent space of three spectral bands (RGB) and standard quantization (0-255), earth observation data is highly non-uniform. Sensors have diverse architectures and observation conditions. Compounding this problem are the complex processing stacks that are added for radiometric and geometric calibration, atmospheric correction, cloud-masking, and other considerations. Two sensors, observing the same location in roughly the same spectral range in the optical spectrum will likely register subtly different readings that may impact our broader image distributions.

Large advances in evaluating EOFMs have been made recently: Lacoste u. a. (2023) Marsocci u. a. (2024), each offer a diversity of tasks, application areas, geographic coverage, and modalities as

[*]Cooresponding Author: rdemilt@sig-gis.com

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

well as thorough evaluation criteria. However, outside of raw performance numbers, the impact of sensor modality on the outputs pretrained models is under-discussed. It is not immediately obvious, for example, that a model pretrained on information collected by a Landsat-8 sensor should even be expected to carry significant symmetry over to samples from the Harmonized Landsat and Sentinel (HLS) Claverie u. a. (2018) reflectance dataset, further into a subset of the same spectral bands collected by a Sentinel-2 satellite, or even beyond to more diverse architectures such as those that capture hyperspectral or extreme high-spatial-resolution imagery. These configurations are non-homogeneous and ofthen times require decades of their own development and engineering.

2 Dataset

To evaluate each model's response to diverse input streams we construct a dataset to dataset for comparison of the same points by many different views. To deepen evaluation it is also valuable to reference our findings to a supervised label which can be used to compare semantics between embedding spaces. To date, no benchmark dataset provides this capability as most datasets include only single modalities or only a single optical modality and SAR components. This is a non-trivial challenge as each modality will have a different acquisition period and therefore some work will need to be done to align these as best as possible.

To meet this challenge we design a collection procedure to generate images with paired samples from each modality. We randomly sample 600 points from the US State of Indiana and draw 224x224 pixel 2-month mosaic composite images at 30-meter spatial resolution for the three optical modalities (and four datasets) examined: HLS, Landsat-8, Landsat-9, Sentinel-2, and our single SAR modality, Sentinel-1. All optical data used is surface reflectance data, and SAR-GRD using the VV and VH polarizations of backscatter readings, in the Interferometric Wide Swath (IW) sensor mode. We utilize the QA pixels of each optical modality to mask images that are more than 20% occluded by clouds. Finally we align the USDA Crop Data Layer as a reference to explore image semantics in our later analysis.

3 Methodology

In order to better evaluate the biases incorporated by the pretraining phase on the representation capabilities of our models we will predominantly restrict ourselves to an unsupervised analysis utilizing encoders with their weights frozen. Many interesting questions can be further explored in the realm of understanding the impact fine-tuning has on the weights and representations of the models used and on the downstream performance implications in specific cases and we provide one exploration of this dimension by examining the performance on the segmentation task with frozen encoders and simple decoder heads.

To perform our evaluations we will utilize the Prithvi Jakubik u. a. (2023) and DOFA Xiong u. a. (2024) foundation models, as representative examples of the broad category of EOFMs with specific characteristics that we will highlight for their utility to our analysis.

The Prithvi model is useful to us because it takes a six spectral band input and was pretrained solely using HLS data. The six input bands from each data stream can be accepted by matching the spectral bands by wavelength categories and discarding the additional bands. This can be seen as archetypal of many EOFM models.

The DOFA model is notable for its use of a dynamic network embedding layer, which utilizes the central wavelengths of the inputs to generate the weights for a convolution-based input layer. The model then processes the image. This allows the model to accept inputs with any band configuration and generate a valid output. However, the model characterizes inputs solely by their central wavelength. The model has also been trained with multiple modalities of data from aerial RGB to satellite data. The flexibility and multimodal training combine to offer a useful counterpoint to the fixed input band configurations which are standard for many models.

3.1 Embedding Space Visualization

The high dimensional nature of the models' latent spaces, 768 for the selected models, makes sample variability difficult to evaluate and impossible to visualize in their native form. Visualization and

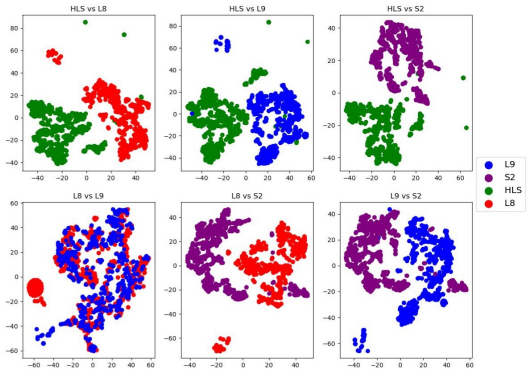

Figure 1: Prithvi TSNE Plots optical modality embeddings

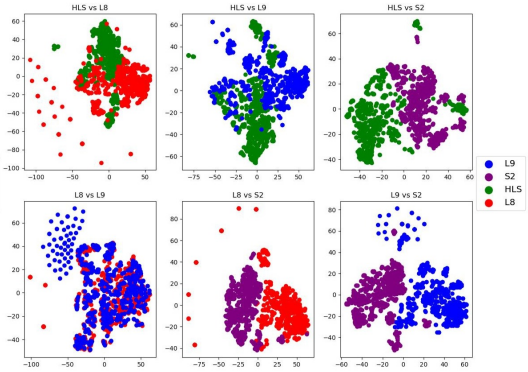

Figure 2: DOFA TSNE Plots optical modality embeddings

Modalities	Prithvi # Matches	Prithvi % Matches	Dofa # Matches	Dofa % Matches
Landsat 8 & Landsat 9	1 / 2 / 4	17.7 / 17.8 / 21.1	1 / 2 / 5	20.60 / 22.1 / 23.9
Landsat 8 & HLS	1 / 2 / 5	22.3 / 24.0 / 26.8	1 / 2 / 5	23.6 / 24.3 / 27.2
Landsat 8 & Sentinel-2	1 / 2 / 4	13.6 / 15.8 / 18.9	1 / 2 / 3	17.4 / 18.6 / 20.4
Landsat 9 & HLS	1 / 3 / 7	28.5 / 30. 5/ 33.1	1 / 3 / 6	29.7 / 28.2 / 29.3
Landsat 9 & Sentinel-2	1 / 2 / 4	16.8 / 18.8 / 22.3	1 / 2 / 5	18.1 / 20.3 / 23.1
HLS & Sentinel-2	1 / 2 / 5	19.0 / 23.1 / 26.7	1 / 2 / 5	21.0 / 21.9 / 25.0

Table 1: Average number of matches (rounded) and average % of local neighborhood agreement between embedded modalities. Each row shows the matches for 5, 10, and 20 nearest neighbors respectively, separated by the "/".

dimensionality reduction techniques such as TSNE van der Maaten und Hinton (2008), offers a perspective on the internal structures of these spaces, and the major axes on which data varies from each other. This makes the technique well suited to our comparative analysis. Ideally, there would not be identifiable clusters based on modality, in either the original heigh-dimensional space or in the low dimensionality reduction, as this would imply that in our high dimensional space data is greatly or principally distributed by factors other than data source. This would be shown best by our embedding spaces being segmentable by semantic features of the image collection such as land cover, crop types, or other differentiating phenomena.

3.2 Local Neighborhood Analysis

The second method for comparing our embedding spaces as we vary input modality is to consider the local neighborhoods of each point based on distance in the embedding space. This is based on the intuitive notion that similarity between data points is a perspective on the overall configuration of the space Duderstadt u. a. (2023); Tavares u. a. (2024) . We can evaluate the similarity of either image patch embeddings or cls tokens. Patch embeddings will evaluate the differentiation in fine grained detail and the cls token should give a summary view of our images. We consider this to be a valuable perspective for two reasons. First, EOFM embeddings of satellite data have been proposed for few-shot and similarity search contexts where the ease of separability and evaluation for embeddings of data has important implications Blumenstiel u. a. (2024). Second, we expect that the subtle distributional differences between our modalities to cause some noise in absolute values of the embedding spaces and at high dimensions these distances are very large in absolute magnitude, by considering neighborhoods we can establish a baseline of what is indicated as similar by a pretrained model and the sensitivity to the sensor architecture.

3.3 Downstream Task Evaluation

In our third evaluation we will present one view on the impact of varying input streams for a downstream task. The evaluation utilizes the Crop Data Layer's labels which are present in the dataset. This segmentation problem is very challenging due to the low quantity of training data relative to the quantity and diversity of crop classes (55 of which are represented in our sample) and

Modality	Random	Prithvi	DOFA
Landsat 8	24.1	57.0	69.3
Landsat 9	24.1	57.6	70.8
HLS	24.1	59.5	71.9
Sentinel-2	24.1	55.5	69.3
Difference	-	4.0	2.6

Table 2: Average Percentage of 10-Nearest Neighborhood with the same mode Crop class for embedded patches in crop data layer labels

EOFM	RF Accuracy	KNN Accuracy
Prithvi	90.7%	85.6%
DOFA	88.7%	91.1%

Table 3: Prediction Accuracy for modality source of patch embed using Random Forest & 5-NN Classifier

the extreme imbalance of classes. This is representative of one of the promises of the EOFM paradigm, the ability to solve problems in limited data environments or with low amounts of training. For training, we utilize the PANGAEA Marsocci u. a. (2024) training protocols and available pretrained model weights. We train with a Fully Convolutional Netowrk (FCN) decoder head as a simple baseline Long u. a. (2015) and the UperNet Xiao u. a. (2018) decoder head used in the PANGAEA benchmark and compare the mean per-pixel accuracy and the mean Intersection over Union when training on each input stream on the test set images after 80 training epochs using the cross entropy loss. In this case we are interested more in the relative performance of our two models when processing differing input streams than in their absolute performance on the segmentation task. As with the previous evaluations, the ideal result would be to observe no variability in the performance between input streams or to observe no consistent variation as this could then be attributable to noise rather than implications of the pretrained weights.

We reiterate here that diverse options for exploring the question of downstream performance of EOFMs can be theorized and explored. These could include fully fine-tuned encoders and diverse decoder training configurations among other consideration. These are important perspectives we hope to see explored by the community but are not the focus of our work. We chooose the outline configuration to highlight the impact of input streams in the frozen encoder or embedding context which we find is a common desired and frequently discussed use case in the community.

4 Results

Using TSNE to visualize the embedding space arrangement in a low dimension, we plot the TSNE transformation of each point and color it by its source modality. We find in Figures 1 & 2 that between most pairs of modalities there are clear clusters of points based on their source modality. The clear separations show that in many cases the minor differences in the distributions of instrument readings are resulting in large variances. It is unsurprising that many of the strongest performing pairings for mixing data of each modality involve the HLS data which is a uniform synthesis of the observations from the other optical data streams and the Landsat pairing which carry nearly identical sensors. The DOFA architecture shows a relatively greater resilience to modality and source change in some cases, compared to Prithvi, but still shows distinct separations in many pairwise cases.

Transitioning to our second evaluation perspective we examine the local neighborhoods of cls tokens for both models. In Table 1 we show another pairwise view of our modalities. Each row lists the rounded number of common neighbors in a 5, 10, and 20-Nearest Neighbors list betweeen two modalities and the average percentage of matching neighbors across the dataset. We find that the percentage of our matching neighbors in local neighborhoods of our cls tokens is only one case greater than 30% on average, meaning that in a context such as similarity search, input modality becomes critically to downstream performance and the determinant of the majority of similar candidates.

Next we examine the neighborhood of the patch embeddings rather than cls tokens. We take the patch embeddings of every patch in our dataset and compare to find the 10-Nearest Neighbors by cosine similarity. As a proxy for semantic similarity at a patch level we calculate the percentage of these neighbors which share a most common class with the original point, using the USDA Crop Data Layer as labels. The average of these over all patches is presented in Table 2. The results show as much as a 4% difference in shared most common class depending on the modality chosen. This provides further evidence that modality change causes not only neighborhood structure to shift but

Model	HLS	S-2	L8	L9
Prithvi+FCN	10.06	8.60	6.34	7.56
Prithvi+Uper	12.61	10.63	10.51	11.85
DOFA+FCN	8.62	7.84	6.04	6.27
DOFA+Uper	12.50	10.29	10.68	12.01

Table 4: Mean image-level IoU on test set for CDL segmentation

Model	HLS	S2	L8	L9
Prithvi+FCN	63.91	59.16	55.08	56.48
Prithvi+Uper	67.71	61.85	60.72	62.42
DOFA+FCN	58.91	54.68	51.73	51.78
DOFA+Uper	66.66	60.48	60.76	63.63

Table 5: Mean pixel-wise Accuracy on test set for CDL segmentation

also creates changes in semantic similarities between local image patches. In a search system this could result in vastly different recall results for a query, or different classification in the few/zero-shot settings. Table 3 shows a complementary perspective wherein two simple classifiers, a RandomForest with 500 estimators and a 5-Nearest Neighbor Classifier, are tasked with predicting the input modality for a patch. We find that with 25% of our patches held out as a test set a simple Random Forest classifier can still get a shocking 90.7% and 88.7% accuracy for Prithvi and DOFA embeddings respectively in predicting which modality a patch embedding originated from, implying it is relatively simple to partition the embedding space of both models into where samples from each modality are placed.

Finally, having established that we can observe clear differences in the arrangement of data points within the embedding space that they are projected onto by our EOFMs relative to the input stream these points originated from, we explore the impact this might have on downstream performance. Again referencing the Crop Data Layer in Table 4 and Table 5 we examine the segmentation capabilities of each model when used as a frozen encoder and training two lightweight decoder heads. We find in this evaluation that both models, across both decoder head configurations have one input stream which is consistently of higher performance. While the exact cause of this preferential performance requires further study, this simple experiment supplements our previous findings by revealing that changing input modality can have implications not merely for similarity but also for performance. We note that the hightest performing input stream for both models is the HLS dataset. For the Prithvi model this was the source of the model's pretraining data. However, for the DOFA model none of these datasets match precisely the pretraining conditions since we are restricting the evaluation to the 6-matching bands between the input streams.

5 Discussion

We find that input modality has a large impact on the outputs of EOFMs. Along these findings we recommend that users of EOFM models pay careful mind to their data collection regime and its alignment with the pretraining data of the chosen model. Combined with the findings in Marsocci u. a. (2024), we propose that it is likely matching spectral bands by wavelength and training on single modalities are insufficient for understanding the complexities of spectral capture technology that extend to subtler underlying distribution shifts than simply matching wavelengths against each other.

Our findings motivate consideration of models based on the sources of their pretraining data and on consideration of diverse sensor modalities for the complexity that clearly underlies them. Additionally we hope this motivates the creation of further highly multimodal and multi-input stream datasets such as our example here to test the cross-modality capabilities of new models.

6 Data Availability Statement

We utilize the model versions available through PANGAEA Marsocci u. a. (2024) to generate embeddings of our multi-modal data. The dataset is available online along with the code used to generate all figures which can be accessed at the following GitHub repository.

References

[Blumenstiel u. a. 2024] BLUMENSTIEL, Benedikt ; MOOR, Viktoria ; KIENZLER, Romeo ; BRUNSCHWILER, Thomas: Multi-Spectral Remote Sensing Image Retrieval Using Geospatial Foundation Models. In: *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing*

Symposium (2024), S. 7286–7291. – URL https://api.semanticscholar.org/CorpusID:268247915

[Claverie u. a. 2018] CLAVERIE, Martin ; JU, Junchang ; MASEK, Jeffrey G. ; DUNGAN, Jennifer L. ; VERMOTE, Eric F. ; ROGER, Jean-Claude ; SKAKUN, Sergii V. ; JUSTICE, Christopher: The Harmonized Landsat and Sentinel-2 surface reflectance data set. In: *Remote Sensing of Environment* 219 (2018), S. 145–161. – URL https://www.sciencedirect.com/science/article/pii/S0034425718304139. – ISSN 0034-4257

[Clay 2024] CLAY, Foundation: *Clay Foundation Model*. https://clay-foundation.github.io/model/index.html. 2024. – [Accessed 11-01-2025]

[Dosovitskiy u. a. 2021] DOSOVITSKIY, Alexey ; BEYER, Lucas ; KOLESNIKOV, Alexander ; WEISSENBORN, Dirk ; ZHAI, Xiaohua ; UNTERTHINER, Thomas ; DEHGHANI, Mostafa ; MINDERER, Matthias ; HEIGOLD, Georg ; GELLY, Sylvain ; USZKOREIT, Jakob ; HOULSBY, Neil: *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. – URL https://arxiv.org/abs/2010.11929

[Duderstadt u. a. 2023] DUDERSTADT, Brandon ; HELM, Hayden S. ; PRIEBE, Carey E.: Comparing Foundation Models using Data Kernels. In: *ArXiv* abs/2305.05126 (2023). – URL https://api.semanticscholar.org/CorpusID:258564434

[Fuller u. a. 2023] FULLER, Anthony ; MILLARD, Koreen ; GREEN, James R.: *CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders*. 2023. – URL https://arxiv.org/abs/2311.00566

[Jakubik u. a. 2023] JAKUBIK, Johannes ; ROY, Sujit ; PHILLIPS, C. E. ; FRACCARO, Paolo ; GODWIN, Denys ; ZADROZNY, Bianca ; SZWARCMAN, Daniela ; GOMES, Carlos ; NYIRJESY, Gabby ; EDWARDS, Blair ; KIMURA, Daiki ; SIMUMBA, Naomi ; CHU, Linsong ; MUKKAVILLI, S. K. ; LAMBHATE, Devyani ; DAS, Kamal ; BANGALORE, Ranjini ; OLIVEIRA, Dario ; MUSZYNSKI, Michal ; ANKUR, Kumar ; RAMASUBRAMANIAN, Muthukumaran ; GURUNG, Iksha ; KHALLAGHI, Sam ; HANXI, Li ; CECIL, Michael ; AHMADI, Maryam ; KORDI, Fatemeh ; ALEMOHAMMAD, Hamed ; MASKEY, Manil ; GANTI, Raghu ; WELDEMARIAM, Kommy ; RAMACHANDRAN, Rahul: *Foundation Models for Generalist Geospatial Artificial Intelligence*. 2023. – URL https://arxiv.org/abs/2310.18660

[Jakubik u. a. 2025] JAKUBIK, Johannes ; YANG, Felix ; BLUMENSTIEL, Benedikt ; SCHEURER, Erik ; SEDONA, Rocco ; MAUROGIOVANNI, Stefano ; BOSMANS, Jente ; DIONELIS, Nikolaos ; MARSOCCI, Valerio ; KOPP, Niklas ; RAMACHANDRAN, Rahul ; FRACCARO, Paolo ; BRUNSCHWILER, Thomas ; CAVALLARO, Gabriele ; BERNABE-MORENO, Juan ; LONGÉPÉ, Nicolas: *TerraMind: Large-Scale Generative Multimodality for Earth Observation*. 2025. – URL https://arxiv.org/abs/2504.11171

[Lacoste u. a. 2023] LACOSTE, Alexandre ; LEHMANN, Nils ; RODRIGUEZ, Pau ; SHERWIN, Evan D. ; KERNER, Hannah ; LÜTJENS, Björn ; IRVIN, Jeremy A. ; DAO, David ; ALEMOHAMMAD, Hamed ; DROUIN, Alexandre ; GUNTURKUN, Mehmet ; HUANG, Gabriel ; VAZQUEZ, David ; NEWMAN, Dava ; BENGIO, Yoshua ; ERMON, Stefano ; ZHU, Xiao X.: *GEO-Bench: Toward Foundation Models for Earth Monitoring*. 2023. – URL https://arxiv.org/abs/2306.03831

[Long u. a. 2015] LONG, Jonathan ; SHELHAMER, Evan ; DARRELL, Trevor: *Fully Convolutional Networks for Semantic Segmentation*. 2015. – URL https://arxiv.org/abs/1411.4038

[Ma u. a. 2019] MA, Lei ; LIU, Yu ; ZHANG, Xueliang ; YE, Yuanxin ; YIN, Gaofei ; JOHNSON, Brian A.: Deep learning in remote sensing applications: A meta-analysis and review. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 152 (2019), S. 166–177. – URL https://www.sciencedirect.com/science/article/pii/S0924271619301108. – ISSN 0924-2716

[van der Maaten und Hinton 2008] MAATEN, Laurens van der ; HINTON, Geoffrey: Visualizing Data using t-SNE. In: *Journal of Machine Learning Research* 9 (2008), Nr. 86, S. 2579–2605. – URL http://jmlr.org/papers/v9/vandermaaten08a.html

[Marsocci u. a. 2024] MARSOCCI, Valerio ; JIA, Yuru ; BELLIER, Georges L. ; KEREKES, David ; ZENG, Liang ; HAFNER, Sebastian ; GERARD, Sebastian ; BRUNE, Eric ; YADAV, Ritu ; SHIBLI, Ali ; FANG, Heng ; BAN, Yifang ; VERGAUWEN, Maarten ; AUDEBERT, Nicolas ; NASCETTI, Andrea: *PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models*. 2024. – URL https://arxiv.org/abs/2412.04204

[Tavares u. a. 2024] TAVARES, Tiago F. ; AYRES, Fabio ; SMARAGDIS, Paris: *Measuring similarity between embedding spaces using induced neighborhood graphs*. 2024. – URL https://arxiv.org/abs/2411.08687

[Wang u. a. 2023] WANG, Yi ; BRAHAM, Nassim Ait A. ; XIONG, Zhitong ; LIU, Chenying ; ALBRECHT, Conrad M. ; ZHU, Xiao X.: *SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation*. 2023. – URL https://arxiv.org/abs/2211.07044

[Xiao u. a. 2018] XIAO, Tete ; LIU, Yingcheng ; ZHOU, Bolei ; JIANG, Yuning ; SUN, Jian: *Unified Perceptual Parsing for Scene Understanding*. 2018. – URL https://arxiv.org/abs/1807.10221

[Xiong u. a. 2024] XIONG, Zhitong ; WANG, Yi ; ZHANG, Fahong ; STEWART, Adam J. ; HANNA, Joëlle ; BORTH, Damian ; PAPOUTSIS, Ioannis ; SAUX, Bertrand L. ; CAMPS-VALLS, Gustau ; ZHU, Xiao X.: *Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation*. 2024. – URL https://arxiv.org/abs/2403.15356