# Landsat-Bench: Datasets and Benchmarks for Landsat Foundation Models

**Isaac Corley** [1] **Lakshay Sharma** [2] **Ruth Crasto** [2]

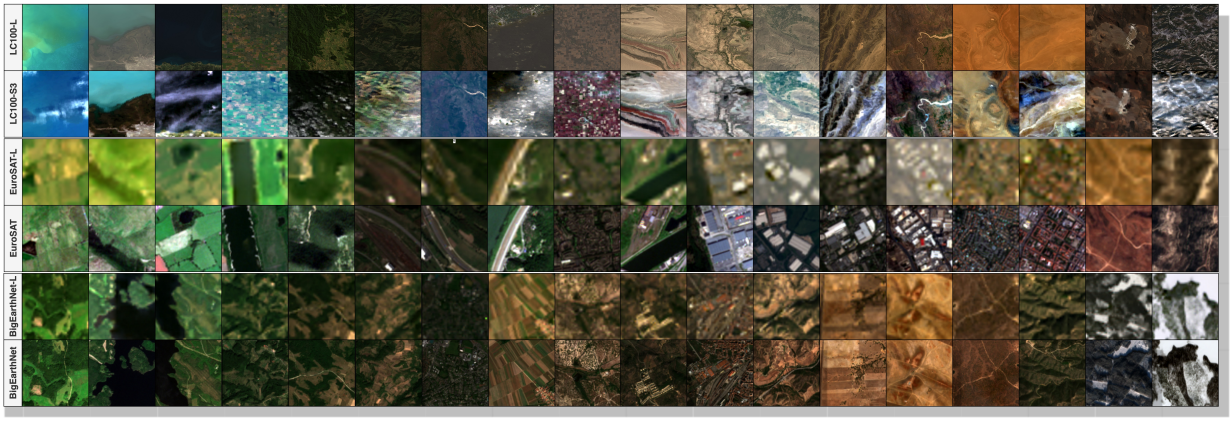


Figure 1: **We introduce `Landsat-Bench`: a new benchmark for Geospatial Foundation Models (GFM) pretrained on Landsat imagery**. Our benchmark contains three Landsat variants (denoted with the suffix *"—L"*.) of existing remote sensing datasets. Samples of each original dataset and our proposed Landsat variants are provided above (from top-to-bottom by row: *LC100-L, LC100-S3, EuroSAT-L, EuroSAT, BigEarthNet-L & BigEarthNet*).

## Abstract

The Landsat program offers over 50 years of globally consistent Earth imagery. However, the lack of benchmarks for this data constrains progress towards Landsat-based Geospatial Foundation Models (GFM). In this paper, we introduce `Landsat-Bench`, a suite of three benchmarks with Landsat imagery that adapt from existing remote sensing datasets – EuroSAT-L, BigEarthNet-L, and LC100-L. We establish baseline and standardized evaluation methods across both common architectures and Landsat foundation models pretrained on the SSL4EO-L dataset. Notably, we provide evidence that SSL4EO-L pretrained GFMs extract better representations for downstream tasks in comparison to ImageNet, including performance gains of +4% OA and +5.1% mAP on EuroSAT-L and BigEarthNet-L. Code and datasets are available at github.com/isaaccorley/landsatbench.

[1]Wherobots [2]Microsoft. Correspondence to: Isaac Corley <isaac@wherobots.com>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

## 1. Introduction

The Landsat program (Roy et al., 2014) has the longest-running continuous record of satellite-based Earth observation, capturing multispectral imagery of every continent with global, temporal, and spectral consistency in comparison to other openly available satellite constellations. This archive consists of petabytes of imagery across several decades, providing an opportunity for building Geospatial Foundation Models (GFM) (Mai et al., 2023) with the potential to generalize across time, space, and environmental conditions. Unlike many modern satellite constellations, Landsat's continuity across multiple generations (Landsat 1 to Landsat 9) makes it uniquely suited for modeling global-scale environmental dynamics, urbanization, and land use change.

Despite being publicly available, Landsat has been largely overlooked in the recent wave of GFM development, which has centered around higher-resolution or more frequently updated sensors like Sentinel-1, Sentinel-2, and NAIP (Wang et al., 2025). This neglect likely stems from a critical gap in benchmarking ability due to the scarcity of curated, standardized benchmarks for Landsat imagery. In this work, we aim to close this gap in favor of motivating more research into Landsat GFMs. Landsat's global consistency, decades of temporal coverage, and relevance to well-studied downstream tasks make it a valuable resource for advanc-

Table 1: **Comparisons of the datasets in `Landsat-Bench` and their original forms.** *ML* and *MT* indicate multi-label and multi-temporal datasets, respectively.

| Dataset | Sensor | Samples | Bands | Res. (m) | Image Size | Classes | ML | MT |
|---|---|---|---|---|---|---|---|---|
| EuroSAT | Sentinel-2 | 27,000 | 13 | 10 | $64 \times 64$ | 10 | × | × |
| EuroSAT-L | Landsat 8 | 27,000 | 7 | 30 | $22 \times 22$ | 10 | × | × |
| BigEarthNet | Sentinel-2 | 590,326 | 12 | 10 | $120 \times 120$ | 19 | ✓ | × |
| BigEarthNet-L | Landsat 8 | 590,326 | 7 | 30 | $40 \times 40$ | 19 | ✓ | × |
| LC100-S3 | Sentinel-3 | 21,692 | 21 | 300 | $96 \times 96$ | 23 | ✓ | ✓ |
| LC100-L | Landsat 8 | 21,690 | 7 | 30 | $960 \times 960$ | 23 | ✓ | ✓ |

ing representation learning in Earth observation. Benchmarks such as Geo-Bench (Lacoste et al., 2023), Copernicus-Bench (Wang et al., 2025), PANGEA (Marsocci et al., 2024), PhilEO Bench (Fibaek et al., 2024), FoMo-Bench (Bountos et al., 2023), Sustainbench (Yeh et al., 2021), and EarthNets (Xiong et al., 2022) have been developed to help researchers empirically evaluate progress in this area.

**Motivation** This work is largely inspired by the SSL4EO-L efforts of (Stewart et al., 2023) to develop a large-scale pretraining dataset of Landsat imagery for building foundation models for use on downstream tasks. Due to the lack of available benchmark datasets for Landsat imagery, the authors were required to modernize two existing cloud detection datasets, L7 Irish (Scaramuzza et al., 2011) and L8 Biome (Foga et al., 2017), and develop two semantic segmentation datasets by aligning Landsat imagery with pixel-level land cover annotations provided by the National Land Cover Database (NLCD) (Dewitz & Survey, 2021) and Cropland Data Layer (CDL) (Boryan et al., 2011).

Prior to SSL4EO-L, SSL4EO-S12 (Wang et al., 2023a) was developed as a large-scale dataset for pretraining self-supervised models on unlabeled satellite imagery from the Sentinel-1 SAR and Sentinel-2 multispectral (MSI) constellations. Due to the wide availability of benchmark datasets for the Sentinel program, the improvement of these foundation models was easily observed across several downstream tasks.

**Related Work** Utilizing existing dataset labels and georeferencing to align with other data products has become accessible due to the existence of catalogs and geospatial data querying tools such as Google Earth Engine (GEE) (Gorelick et al., 2017) and Microsoft's Planetary Computer (Microsoft, 2022). Wang et al. (2023b) developed the EuroSAT-SAR dataset, a Sentinel-1 version of the EuroSAT benchmark dataset (Helber et al., 2019), using GEE to evaluate SAR foundation models. Cong et al. (2022) created a Sentinel-2 version of the fMoW dataset (Christie et al., 2018) to evaluate their SatMAE architecture pretrained on Sentinel-2 imagery.

**Contributions** In summary, there has recently been significant progress in developing large pretraining datasets for several satellite constellations; however, additional benchmark datasets for quantifying downstream task performance of satellite-specific foundation models are still needed. In this paper, we look to further close this gap. We detail our contributions below:

- We introduce the first iteration of `Landsat-Bench` consisting of three benchmark datasets for evaluating geospatial foundation models pretrained on Landsat imagery: **LC100-L**, **EuroSAT-L**, and **BigEarthNet-L**.
- We provide baselines for our benchmarks of standard deep learning architectures and the SSL4EO-L line of geospatial foundation models.

## 2. Datasets

We introduce three new datasets for evaluating machine learning models trained on Landsat imagery – *LC100-L*, *EuroSAT-L*, and *BigEarthNet-L*. We use existing labeled remote sensing datasets commonly used in benchmarking foundation models trained on satellite imagery, namely the LC100-S3 (Wang et al., 2025), EuroSAT (Helber et al., 2019), and BigEarthNet (Sumbul et al., 2019) datasets. We utilize the original dataset implementations from the `torchgeo` library (Stewart et al., 2022). Descriptions of the datasets and our Landsat variants are provided in Table 1. Below are details of each original dataset.

**LC100** The LC100 dataset is a multi-label classification dataset which is part of the CopernicusBench (Wang et al., 2025) benchmark suite. The dataset contains $96 \times 96$ 300m spatial resolution images of Sentinel-3 imagery with 21 multispectral bands. The image classification labels contain 23 categories derived from the 100m land cover product from the Copernicus Global Land Service (CGLS) (Buchhorn et al., 2020). We utilize the dataset splits provided with the data.

**EuroSAT** The EuroSAT dataset (Helber et al., 2019) is a land cover classification dataset of patches extracted

Table 2: **Baseline evaluation of 9 pretrained models on `Landsat-Bench`**. We report both KNN (k=5) and Linear Probe (LP) results on extracted features of each method. For EuroSAT-L classification we report overall accuracy (OA). For BigEarthNet-L and LC100-L multilabel classification we report (micro) mean average precision (mAP). * denotes the model was pretrained on the SSL4EO-L dataset using the weights from (Stewart et al., 2023) made available in `torchgeo`.

| | | | ResNet-18 (11.7M) | | | ResNet-50 (25.6M) | | | ViT-S (22.1M) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | **Dataset** | **Zonal Stats** | **ImageNet** | **MoCo v2*** | **SimCLR*** | **ImageNet** | **MoCo v2*** | **SimCLR*** | **ImageNet** | **MoCo v2*** | **SimCLR*** |
| **KNN** | EuroSAT-L | 78.9 | 76.2 | 79.0 | 69.8 | 76.4 | 75.1 | 72.6 | 79.4 | 81.6 | **83.4** |
| | BigEarthNet-L | 64.5 | 57.5 | 61.4 | 51.4 | 58.3 | 59.3 | 53.9 | 61.9 | 64.5 | **67.0** |
| | LC100-L | 56.3 | 56.5 | 56.9 | **57.7** | 56.6 | 56.2 | 56.8 | 56.5 | 56.0 | 55.8 |
| **LP** | EuroSAT-L | 74.1 | 81.0 | 82.8 | 68.5 | 79.0 | 82.5 | 73.5 | 86.4 | **88.8** | 77.6 |
| | BigEarthNet-L | 60.9 | 70.2 | 74.0 | 60.3 | 71.9 | 76.1 | 63.1 | 74.9 | **77.5** | 68.3 |
| | LC100-L | 62.3 | 62.1 | 61.6 | **63.8** | 60.6 | 60.5 | 62.1 | 60.5 | 59.9 | 63.4 |

from multispectral Sentinel-2 imagery. The dataset contains 27,000 64 × 64 10m spatial resolution images with 13 bands and labels for 10 land cover categories. We use the dataset splits defined in (Neumann et al., 2019).

**BigEarthNet** The BigEarthNet dataset (Sumbul et al., 2019) is a multi-label land cover classification dataset of patches extracted from multispectral Sentinel-2 imagery. The dataset contains 590,326 120 × 120 10m spatial resolution images with 12 multispectral bands and labels for 19 land cover categories. We use the splits provided with the dataset and defined in (Sumbul et al., 2019).

**Data Acquisition** We extract the timestamps and geospatial bounds of the imagery of each dataset. We then use this metadata to download corresponding imagery from the Landsat Collection 2 Level-2 dataset using Microsoft's Planetary Computer (Microsoft, 2022). We specifically download imagery from the Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) (Engebretson, 2020) containing the following 7 multispectral bands – Coastal/Aerosol, Blue, Green, Red, NIR, SWIR-1, and SWIR-2 – which correspond to the following band IDs – (`SR_B1`, `SR_B2`, `SR_B3`, `SR_B4`, `SR_B5`, `SR_B6`, `SR_B7`). All bands are provided with a spatial resolution of 30 m/pixel.

To ensure imagery contains as few invalid pixels as possible, we constrain the extracted Landsat patches to meet a minimum valid pixel ratio of 80%.

While the BigEarthNet and LC100 imagery have associated timestamps, the EuroSAT dataset does not. Therefore, we initially select a 6-month range on the year of dataset release between $2018/03/01 - 2018/08/31$. For the BigEarthNet and LC100 imagery, we select a range of $\pm 30$ days from the original timestamp. Due to a few images not being available within the 30-day range, we iteratively increase the range by an additional 30 days until all images are downloaded.

Although the ideal case is to download cloud-free imagery as close to the original timestamp as possible, we observe that a small number of samples require us to lessen the constraints. We initially query for imagery with $\leq 10\%$ cloud cover and with the above timestamp rules. We then sample the image closest to the original timestamp with the least amount of clouds. We log the samples which do not have available imagery and progressively lessen the constraints to $\leq 20\%$ cloud cover and increase the delta of months from the timestamp by 1 until we have collected all remaining samples.

## 3. Benchmarks

**Baseline Models** We evaluate a selection of common convolutional and transformer-based architectures – **ResNet-18, ResNet-50, and ViT-S/16**. We compare the performance of model weights pretrained for Landsat imagery vs. a baseline of ImageNet pretraining using the implementations provided in the `torchgeo` library (Stewart et al., 2022). We also include a naive baseline to our experiments, referred to as **Zonal Stats** where we compute a handcrafted feature embedding for each patch by summarizing each spectral band using per-band mean, standard deviation, minimum, and maximum values and concatenating them into a fixed-length feature vector and applying standard normalization to scale the features to similar ranges. This is analogous to zonal statistics in remote sensing, where aggregate statistics are computed over spatial zones. The following pretraining methods are detailed below:

**ImageNet** We use ImageNet pretrained weights as a baseline, in particular we use the default weights and the first convolutional layer adaptive initialization scheme from the `timm` library (Wightman, 2019) which repeats the RGB weights for additional channels, in our case 7 Landsat multispectral bands.

**SSL4EO-L MoCo v2** We utilize weights from (Stewart et al., 2023) pretrained using the MoCo v2 (Chen et al., 2020b) self-supervised learning method on the SSL4EO-L dataset.

Table 3: **Comparison of linear probing results on `Landsat-Bench` vs. current state-of-the-art on the original datasets.** For EuroSAT we report overall accuracy (OA). For BigEarthNet and LC100 we report micro mean average precision (mAP).

| Dataset | Method | Params (M) | Metric | Performance |
|---|---|---|---|---|
| EuroSAT | Copernicus-FM ViT-B/16 (Wang et al., 2025) | 139.5 | OA | 97.9 |
| EuroSAT-L | SSL4EO-L MoCo v2 ViT-S/16 (Stewart et al., 2023) | 22.1 | OA | 88.8 |
| BigEarthNet | SoftCon ViT-B/14 (Wang et al., 2024) | 86.6 | mAP | 86.1 |
| BigEarthNet-L | SSL4EO-L MoCo v2 ViT-S/16 (Stewart et al., 2023) | 22.1 | mAP | 77.5 |
| LC100-S3 | Copernicus-FM ViT-B/16 (Wang et al., 2025) | 139.5 | mAP | 93.3 |
| LC100-L | SSL4EO-L SimCLR ResNet-18 (Stewart et al., 2023) | 11.7 | mAP | 63.8 |

**SSL4EO-L SimCLR** We utilize weights from (Stewart et al., 2023) pretrained using the SimCLR (Chen et al., 2020a) self-supervised learning method on the SSL4EO-L dataset.

**Experiments** To evaluate existing foundation models' ability to extract useful feature representations for downstream applications, we perform both K-Nearest Neighbor (KNN) (k=5) and Linear Probing (LP) evaluations. To avoid conflating any performance gains with custom implementations of evaluation pipelines, we keep our benchmarks simple and utilize the KNN and Logistic Regression implementations from `scikit-learn` (Pedregosa et al., 2011) to operate directly on extracted image embeddings. For Linear Probing of the multi-label datasets (BigEarthNet-L and LC100-L) we employ a One-vs-Rest classifier using a Stochastic Average Gradient (SAG) solver and a maximum of 1,000 iterations. For preprocessing, we use the resize and normalization strategy recommended by (Corley et al., 2024).

## 4. Discussion & Future Work

**Benchmark Results** The results in Table 2 provide evidence overall that self-supervised learning methods pretrained on Landsat imagery outperform ImageNet pretraining by a significant margin, e.g. $+4\%$ OA and $+5.1\%$ mAP for the EuroSAT-L and BigEarthNet-L KNN results, respectively. However, selecting which method appears to depend on the task. MoCo v2 tends to significantly outperform SimCLR for linear probing, while the opposite is true for KNN evaluations. For LC100-L, KNN results across all models remained similar; only after linear probing did the differences in model performance become apparent.

**Computational Efficiency** Landsat's 30m resolution offers a balance between detail and efficiency, avoiding potential increased data and compute costs of 10m sensors like Sentinel-2. For tasks like land-cover classification or change detection, this trade-off can enable faster analysis with minimal loss in performance. As a result, Landsat GFMs can be suitable for resource-constrained geospatial tasks.

For context, assume we want to perform global inference with a ResNet-50 on a NVIDIA V100 GPU with a patch size of $224 \times 224$ at a throughput of 500 images/sec. For the Earth's total land area of $1.49 \times 10^{14}$ m$^2$, the inference time to process all global patches, assuming non-overlapping patches, using Equation 1 results in approximately **1.83 hours** for Landsat-8 and **16.5 hours** for Sentinel-2 (an increase of **9**$\times$ as expected due to the 9$\times$ increase in resolution).

$$\mathrm{T\ (s)} = \frac{\mathrm{Area\ (m^2)}}{\mathrm{Res^2\ (m^2/px^2)} \cdot \mathrm{Patch\ Size^2\ (px^2)} \cdot \mathrm{Throughput\ (img/s)}} \tag{1}$$

**Sensor Ensembling** Training models for the same tasks for both Landsat and Sentinel constellations can benefit from increased revisit rates. By merging Sentinel's revisit rate of 5 days with Landsat 8 & 9's rate of 8 days, using an ensemble of multi-sensor models increases the rate at which time-sensitive applications can be performed, such as wildfire (de Almeida Pereira et al., 2021) and deforestation (Watch, 2002) monitoring.

**Thermal Infrared Applications** While Landsat provides a lower resolution for optical bands than Sentinel-2, its Thermal Infrared Sensor (TIRS) provides 100m thermal surface temperature measurements which are unavailable in the Sentinel-2 constellation. While Sentinel-3 captures thermal infrared, it comes at a 10x lower resolution of 1,000m. This enables applications where thermal monitoring is essential such as urban heat island detection (Keeratikasikorn & Bonafoni, 2018) and crop water stress assessment (Ahmad et al., 2021).

**Benchmark Difficulty** While this can also be seen as a positive, we note there is a significant drop in performance compared to the original datasets – as seen in Table 3. Most notably, LC100-L performance drops from 93.8 to 63.8 mAP for the current best models. This can be attributed to LC100 originally being a Sentinel-3 derived dataset with 21 multispectral bands, potentially containing spectral features for class separability that the Landsat 8 OLI 7 bands do not provide.

## References


Ahmad, U., Alvino, A., and Marino, S. A review of crop water stress assessment using remote sensing. *Remote Sensing*, 13(20):4155, 2021.

Boryan, C., Yang, Z., Mueller, R., and Craig, M. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

Bountos, N. I., Ouaknine, A., and Rolnick, D. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*, 2023.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.-E., Herold, M., and Fritz, S. Copernicus global land service: land cover 100 m: collection 3: epoch 2019: Globe. *Zenodo*, 2020. doi: 10.5281/zenodo.3939050. URL https://doi.org/10.5281/zenodo.3939050.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

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

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Corley, I., Robinson, C., Dodhia, R., Ferres, J. M. L., and Najafirad, P. Revisiting pre-trained remote sensing model benchmarks: resizing and normalization matters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3162–3172, 2024.

de Almeida Pereira, G. H., Fusioka, A. M., Nassu, B. T., and Minetto, R. Active fire detection in landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178: 171–186, 2021.

Dewitz, J. and Survey, U. G. National Land Cover Database (NLCD) 2019 products (ver. 2.0). U.S. Geological Survey data release, June 2021. URL https://doi.org/10.5066/P9KZCM54.

Engebretson, C. Landsat 8–9 Operational Land Imager (OLI) - Thermal Infrared Sensor (TIRS) Collection 2 Level 1 (L1) Data Format Control Book (DFCB). Technical report, Department of the Interior, U.S. Geological Survey, September 2020. LSDS-1822.

Fibaek, C., Camilleri, L., Luyts, A., Dionelis, N., and Le Saux, B. Phileo bench: Evaluating geo-spatial foundation models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2739–2744. IEEE, 2024.

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, R. D., Beckmann, T., Schmidt, G. L., Dwyer, J. L., Hughes, M. J., and Laue, B. Cloud detection algorithm comparison and validation for operational landsat data products. *Remote sensing of environment*, 194:379–390, 2017.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL https://doi.org/10.1016/j.rse.2017.06.031.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Keeratikasikorn, C. and Bonafoni, S. Urban heat island analysis over the land use zoning plan of bangkok by means of landsat 8 imagery. *Remote Sensing*, 10(3):440, 2018.

Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

Marsocci, V., Jia, Y., Bellier, G. L., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

Microsoft. microsoft/planetarycomputer: October 2022, October 2022. URL https://doi.org/10.5281/zenodo.7261897.

Neumann, M., Pinto, A. S., Zhai, X., and Houlsby, N. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145:154–172, 2014.

Scaramuzza, P. L., Bouchard, M. A., and Dwyer, J. L. Development of the landsat data continuity mission cloud-cover assessment algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1140–1154, 2011.

Stewart, A. J., Robinson, C., Corley, I. A., Ortiz, A., Ferres, J. M. L., and Banerjee, A. Torchgeo: deep learning with geospatial data. In *Proceedings of the 30th international conference on advances in geographic information systems*, pp. 1–12, 2022.

Stewart, A. J., Lehmann, N., Corley, I. A., Wang, Y., Chang, Y.-C., Braham, N. A. A., Sehgal, S., Robinson, C., and Banerjee, A. Ssl4eo-l: Datasets and foundation models for landsat imagery. *arXiv preprint arXiv:2306.09424*, 2023.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE, 2019.

Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X. 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, 2023a.

Wang, Y., Hernández, H. H., Albrecht, C. M., and Zhu, X. X. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint arXiv:2310.18653*, 2023b.

Wang, Y., Albrecht, C. M., and Zhu, X. X. Multi-label guided soft contrastive learning for efficient earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Wang, Y., Xiong, Z., Liu, C., Stewart, A. J., Dujardin, T., Bountos, N. I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., et al. Towards a unified copernicus foundation model for earth vision. *arXiv preprint arXiv:2503.11849*, 2025.

Watch, G. F. Global forest watch. *World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002)*, 2002.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Xiong, Z., Zhang, F., Wang, Y., Shi, Y., and Zhu, X. X. Earthnets: Empowering ai in earth observation. *arXiv preprint arXiv:2210.04936*, 2022.

Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D. B., and Ermon, S. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.