MMEarth-Bench: Global Model Adaptation via Multimodal Test-Time Training

Lucia Gordon[1,2], Serge Belongie[2], Christian Igel[2], and Nico Lang[2]

[1] Harvard University, USA
[2] University of Copenhagen, Denmark
luciagordon@g.harvard.edu, nila@di.ku.dk

Abstract. Recent research in geospatial machine learning demonstrates that models pretrained with self-supervised learning on Earth observation data can perform well on downstream tasks with limited labeled data. However, most benchmark datasets have few data modalities and poor global representation, limiting the ability to evaluate multimodal pretrained models at global scales. In order to fill this gap, we introduce MMEarth-Bench, a collection of five new environmental tasks with 12 modalities, globally distributed data, and both random and geographic test splits. We benchmark a diverse set of pretrained models and find that while (multimodal) pretraining tends to improve model robustness in limited data settings, geographic generalization abilities remain poor. Moreover, a simple randomly initialized multimodal model is competitive given enough labeled data. Although data is abundant, models can currently only make use of the modalities on which they were pretrained. To solve this problem, we propose using all the modalities available at test time as auxiliary tasks for test-time adaptation. Our model-agnostic method for test-time training with multimodal reconstruction (TTT-MMR) can improve performance across all models and tasks on both test splits. Furthermore, geographic batching leads to a good trade-off between regularization and specialization during TTT, which is especially beneficial for long-tail distributions. Our dataset, code, and visualization tool are linked on the project page: lgordon99.github.io/mmearth-bench.

Keywords: Earth observation · Multimodality · Test-time adaptation

1 Introduction

Earth observation (EO) data helps us monitor our planet's health, respond to natural disasters, and tackle societal challenges such as the Sustainable Development Goals (SDGs) [38]. A multitude of sensors provide unprecedented data about the planet but also far exceed what most data processing pipelines can handle. To realize the full potential of EO data, raw observations, such as optical satellite images, must be interpreted at both local *and* global scales. Deep learning seems like a natural approach to analyzing copious unlabeled data, and supervised learning has shown great success in the case of abundant reference data [63], but we identify *three grand challenges* yet to be overcome (see Fig. 1).

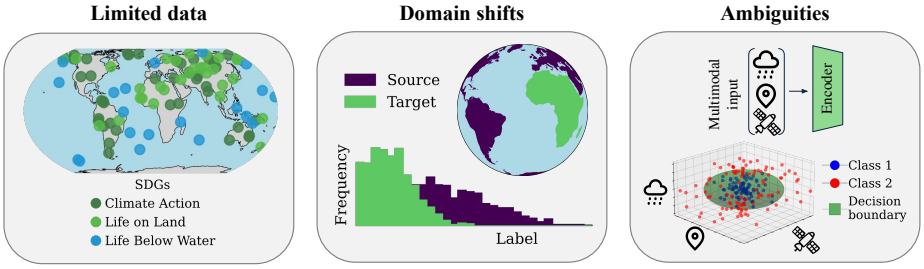

Fig. 1: Self-supervised multimodal pretraining promises to overcome three grand challenges in Earth observation. Crucial applications have to rely on *limited and sparse* and *geographically biased* training data. Furthermore, the *ambiguities* inherent to modeling biophysical quantities with remotely sensed data may be resolved by models conditioned on *multiple modalities*.

The first long-standing challenge has been the sparsity and limited amount of labeled data, as is often the case in crucial applications where reference data is collected through field measurements [63]. The second challenge is geographic generalization, *i.e.* the domain shift that occurs when using a model trained on one geographic region to make predictions on another [31]. Lastly, the third main challenge is to minimize the ambiguities inherent to estimating biophysical quantities from observational, especially unimodal, EO data [40]. The research community working at the intersection of machine learning and EO has identified self-supervised learning (SSL) as a promising way forward [1,23,35,51]. Pretraining models with global unlabeled data promises to facilitate model adaptation to tasks with limited labels and ameliorate performance drops in geographic regions without reference data. Because EO data is georeferenced, the multitude of sensors and satellites means that there are myriad data modalities that can be automatically aligned [35]. This is an opportunity for multimodal data fusion to resolve ambiguities in model predictions [39, 42, 45, 50]. Hence, EO is relevant to the computer vision community for advancing multimodal representation learning [33]. While existing EO benchmarks [7, 16, 25, 31, 53] have helped to develop and evaluate pretrained geospatial models, they limit the evaluation of progress towards overcoming these three grand challenges, as most of their downstream tasks are i) regional rather than global, ii) not evaluated on geographic splits with an entire held-out region, and iii) only provide few modalities per task.

In this work we introduce MMEarth-Bench, a multimodal EO benchmark dataset containing five global tasks designed to evaluate pretrained geospatial models in these three challenging yet crucial settings. As in prior work, we use the term "modality" to describe sensor measurements, metadata, and derived products [35]. We evaluate 8 pretrained models on our benchmark and find that multimodal pretraining does improve performance in low-shot settings and under geographic shifts. However, we show that given enough labeled data, training a multimodal model from scratch is a strong baseline. As is the case with our evaluation and more generally, there is a mismatch between the *input modalities*

on which a model was pretrained and the *task modalities* available at inference time [31]. The standard practice is simply to use only the task modalities that the pretrained model can take as input. To address this inefficiency, we introduce a method for *using all available modalities at test time, whether or not they are compatible with the pretrained model*. Complementary to using multimodal data for self-supervised pretraining, we propose using the modalities as reconstruction tasks in the regime of test-time adaptation, which uses unlabeled test data to adapt the model at test time. We formulate our proposed method for Test-Time Training with MultiModal Reconstruction (TTT-MMR), along with a second variant using geographic batching (TTT-MMR-Geo), which leads to a good trade-off between specialization and regularization, a crucial aspect in TTT [19, 32]. We summarize our major contributions:

1. **MMEarth-Bench dataset:** We introduce five multimodal environmental monitoring tasks for benchmarking pretrained models. Each task consists of 12 aligned modalities and provides globally distributed data with both random and geographic test splits.
2. **Benchmarking:** We benchmark 8 recent pretrained models covering a range of architectures and including both uni- and multimodal models. We find that all models still struggle to generalize geographically and that training a multimodal model from scratch is competitive with global pretraining.
3. **Method:** To advance model adaptation when labels are limited and geographically biased, we propose repurposing multimodal pretext tasks for test-time adaptation, using modality reconstruction losses as a test-time adaptation signal. Our model-agnostic approach yields performance improvements across all models and tasks on both test splits.

2 Related Work

2.1 Low-shot learning

Pretrained models. One strategy to facilitate low-shot learning is self-supervised pretraining. In contrast to vision foundation models trained on web data [12, 46], EO models are pretrained on large amounts of unlabeled remotely sensed data. Hong *et al*. [21] provide an overview of the myriad foundation models in remote sensing. While some geospatial models are still pretrained on RGB imagery [2,27,30,41,46], more recent unimodal models include additional spectral bands [3, 10, 35]. We benchmark various pretrained models for global adaptation with limited and geographically biased labels, as found in real applications.

Multimodal fusion. Subsequently, additional work aimed to learn richer geospatial representations by pretraining on multimodal data [1, 23, 53, 56] or training multimodal models from scratch [40]. We evaluate the benefits of modality fusion in *pretrained* models under geographic shifts. Moreover, whereas multimodal models use multiple modalities as input, our method uses them as *targets*.

Joint training (JT). Joint training trains a model simultaneously on both the main task and auxiliary tasks in order to improve robustness on the main

task, especially when training data is limited [20,47,49]. The model has a shared encoder and separate heads for the main task and self-supervised tasks [47]. This method only uses multimodal data to generate reconstruction losses at train time. In contrast, our method also leverages the reconstruction losses as an *adaptation signal at test time*. We use JT as a baseline in our TTT experiments.

2.2 Domain adaptation

Unsupervised domain adaptation (UDA). Unsupervised domain adaptation uses labeled data from the source domain and unlabeled data from the target domain to adapt to the target domain [29]. UDA-SS [48] extends JT by including unlabeled target data during the training process to facilitate the encoder's adaptation to the target domain. Like many common UDA methods, this is primarily designed for covariate shifts across domains [28, 29]. The geographic generalization problem differs, as we also encounter *label* distribution shifts. Scheibenreif *et al*. [43] insert adapters into a pretrained encoder and train them with SSL using target domain modalities before finetuning the model with target domain labels. In contrast, our problem setup assumes that we neither have access to labeled nor unlabeled data from the target domain during training. Our model-agnostic approach also avoids modifying the encoder architecture.

Test-time adaptation (TTA). Unlike traditional UDA which uses target distribution data during training, test-time adaptation, or equivalently test-time training (TTT), updates the model during testing to address distribution shifts [17,49]. Following JT, rather than discarding the self-supervised task head [48], TTT uses this head at test time to calculate a loss for adapting the encoder before making a final prediction [17,49]. Past works have used rotation prediction [49] or masked reconstruction [17] as auxiliary tasks. To avoid having to define a self-supervised auxiliary task, TENT [52] uses the prediction entropy as an adaptation signal. This makes it a natural approach for classification tasks but not for regression, which is common for biophysical variables (4 of our 5 tasks). Unlike TENT, which updates the encoder's batch norm statistics, our proposed TTT method works with any architecture and loss. This is especially relevant as many pretrained EO models use a transformer architecture without batch norm layers. Sparse or coarse labels have also been used to produce a TTA signal [60]. In the spirit of using weak labels for TTT, we use the data modalities available at test time as reconstruction targets. While TTA has been used with remote sensing imagery to adapt to distribution shifts in street-level RGB imagery [54] and guide robots' search for targets in an environment [32], we adapt pretrained models to *diverse* environmental tasks and domains at *global* scales.

Regularization in TTA. Regularization is crucial for TTA [19, 32], and simply using batched test data is a regularization method, as it results in less noisy gradients [36,62]. Test batches can be comprised of multiple augmentations of a single image [49] or just be random samples of the test data [17]. We propose geographic batching as a compromise between regularization and specialization. Updating only the batch norm parameters [52] or FiLM layers [37,60] are commonly used regularization approaches. In model-agnostic fashion, our method's

adaptation signal uses 12 modality reconstruction losses, as opposed to just one, for regularization. Each individual modality's gradient may be noisy, but averaging over them reduces noise. In this way, we avoid needing to insert batch norm or FiLM layers. The S4T method [24] uses a vision transformer as a task behavior synchronizer to lessen degradations in model performance due to the ideal number of TTT iterations differing among multiple auxiliary tasks. We achieve regularization with a lightweight linear task modality decoder.

Multimodal TTA. Shin *et al.* [44] study multimodal TTA for 3D semantic segmentation by producing pseudo-labels for use as a self-training signal at test time given RGB images and LiDAR point clouds. Like TENT, their method updates batch norm parameters in the architecture. READ [57] addresses the scenario where modalities vary in their reliability when subject to domain shifts. They encode each modality with a transformer and modulate attention-based fusion layers at test time. Unlike READ, our approach does not require encoding the auxiliary modalities, instead solely using them as reconstruction targets.

2.3 Existing benchmark datasets

Overview. Past work in benchmark datasets for pretrained EO models has guided progress in model development and pretraining strategies. We summarize the most related benchmarks in Tab. 1. *SustainBench* [59] is a collection of 15 tasks designed to track progress in 7 sustainable development goals, including no poverty, quality education, and climate action. The modalities vary by task and can include imagery from Landsat, Sentinel-2 (S2), MODIS, NAIP, as well as geolocation and time. *GEO-Bench* [25] contains 12 tasks ranging from building and solar panel classification to tree segmentation. The modalities vary by task and can include S2, Sentinel-1 (S1), Landsat-8, RGB, hyperspectral imagery, or elevation data. None of the tasks are globally distributed. *FoMo-Bench* [7] contains 15 datasets for forest monitoring, where each task has up to 4 modalities such as S2, S1, LiDAR, elevation, or meteorological data. *PhilEO Bench* [16] contains task data for building density estimation, road segmentation, and landcover classification. All tasks have S2 as the single input modality. *Copernicus-Bench* [53] is a benchmark focusing on cross-modal model evaluation that includes 15 unimodal tasks including land cover, biomass, and air quality.

Table 1: Comparison of benchmark datasets for EO (vision) models. "M": number of modalities per task. "✓/✗": a property holds for a subset of tasks.

Benchmark	Domain	M	Global	Climate
Copernicus-Bench [53]	Mixed	1	✓/✗	✓/✗
PANGAEA [31]	Mixed	1-3	✓/✗	✗
PhilEO Bench [16]	Built-up	1	✓	✗
FoMo-Bench [7]	Forests	1-4	✓/✗	✓/✗
GEO-Bench [25]	Mixed	1-3	✗	✗
SustainBench [59]	SDGs	1-10	✓/✗	✗
MMEarth-Bench (ours)	Nature	12	✓	✓

The input modality varies across tasks and can include sensors such as Sentinel-1, -2, -3, or -5P. PANGAEA [31] is a collection of existing benchmark datasets, each of which has up to 3 modalities such as S2, S1, Planet, or Maxar.

Limitations. None of the existing benchmark datasets share the same multiple modalities across all downstream tasks, and the vast majority of tasks contain no more than 3 modalities. This limits the evaluation of multimodal pretrained models, since likely not all of their modalities are available at test time. Furthermore, the vast majority of downstream tasks are limited to a single geographic region. Even if a benchmark contains tasks covering different regions, this does not make it possible to determine whether a pretrained model is able to generalize from one region to another. While PANGAEA explicitly mentions a geographic test split for one task and SustainBench includes geographic splits for several tasks, benchmarks rarely explicitly state whether a test split represents an entire held-out region or simply avoids overlap with the training data. This is likely due to a combination of geographic generalization not being prioritized and many benchmarks collecting existing datasets and using their data splits. The degree to which a geographic test split can be a distribution shift is also limited by the geographic spread of the data in the first place. Additionally, few downstream tasks include climate data as a modality, limiting development in fusion methods for optical EO and climate data. While there are several downstream tasks focusing on the natural world, many tasks focus on the human-nature interface and the built-up world. Moreover, the same tasks appear in multiple benchmarks, most of which include preexisting datasets. We contribute five novel, global datasets for evaluating pretrained models on biomass, soil property, and species occurrence prediction. Our tasks focus on the natural world with 12 shared modalities across imagery, map products, and climate variables. These design choices should facilitate future development of multimodal models.

3 Dataset

Overview. MMEarth-Bench is comprised of five new datasets, summarized in Tab. 2, for environmental tasks: aboveground biomass, soil nitrogen (N), soil organic carbon (OC), soil pH, and species occurrence. Each single-timestamp dataset includes 12 modalities along with task data for each 128×128-pixel tile. Every pixel represents 10m on the ground, so each tile spans an area of $\approx 1.6\ \text{km}^2$. All five tasks share the same 12 modalities (see Tab. 3) present in the MMEarth pretraining dataset [35]. Within each task, we ensure no overlap among tiles. Note that the MMEarth-Bench datasets are not designed for producing the next best task-specific model or for time-series modeling. Rather, we design our datasets for systematic evaluation of pretrained models that aim to generalize to various downstream tasks with limited data under geographic shifts.

Splits. Each task's tiles are split into train, validation, random test, and geographic test sets (see Fig. 2). The tiles in Africa define the geographic test set, as it is often underrepresented in labeled data derived from field measurements, as is the case for our soil tasks [11, 55]. In practice, models finetuned with non-

Table 2: MMEarth-Bench tasks.

Task	# Tiles	Unit	Scale	Type	License
Biomass	18,393	Mg/ha	Pixel-level	Regression	CC BY
Soil Nitrogen	5,643	g/kg	Tile-level	Regression	CC BY-NC
Soil Organic Carbon	7,982	g/kg	Tile-level	Regression	CC BY-NC
Soil pH	8,508	Unitless	Tile-level	Regression	CC BY-NC
Species	36,410	Occurrence	Tile-level	Multi-label classification	Terms of Use

Table 3: MMEarth-Bench modalities.

Pixel-level

Modality	Bands / Variables
Sentinel-2 (S2)	B1–B8, B8A, B9, B11, B12
Sentinel-1 (S1)	(Asc., Desc.) × (VV, VH, HH, HV)
ASTER GDEM	Elevation, slope
ETH Canopy Height	Height, uncertainty
Dynamic World	Landcover (9 categories)
ESA WorldCover	Landcover (11 categories)

Tile-level

Modality	Bands / Variables
Precipitation	Last month, month, year
Temperature	(Last month, month, year) × (max, mean, min)
Geolocation	Longitude, latitude
S2 Date	Date
Biome	Biome (14 categories)
Ecoregion	Ecoregion (846 categories)

Africa data might be used to make predictions in Africa, making it especially relevant to understand how well this generalization works. The remaining tiles are randomly split into train/validation/random test with ratios 70%/15%/15%. We include a random and a geographic test set to evaluate model performance both in- and out-of-distribution. To allow for evaluating models in very low data regimes, we also provide train sets containing 50% and 5% of the training tiles, selected randomly but where the 5% train set is a subset of the 50% train set.

Modalities. All tiles include 12 modalities, six of which are pixel-level (S2, S1, ASTER GDEM [34], ETH Global Canopy Height [26], Dynamic World [8] and ESA WorldCover [61]) and six of which are tile-level (ERA5 precipitation and temperature, geolocation, Sentinel-2 date, biome and ecoregion [13]). Eight of the modalities are continuous-valued and four are categorical-valued. All of the modality data is accessed through Google Earth Engine (GEE) [18].

Tasks. Our biomass dataset is sourced from aboveground biomass estimates from NASA's 2020 GEDI mission [14]. We sample measurements with GEE such that our tiles are approximately balanced across biomes. We obtain our soil data from the WoSIS December 2023 snapshot [4]. We select soil N, soil OC, and soil pH due to their environmental and ecological significance. Nitrogen is crucial for plant growth, organic carbon is an indicator of soil quality, and pH has a strong influence on biogeochemical processes in soil [58]. For each soil property we aggregate data across time within the depth range 0-5 cm. Biomass and the soil properties are regression tasks. We extract our species range data from the IUCN Red List's terrestrial mammal range polygons [22]. We filter by species whose historic range covers at least 6,000 km^2 both in and outside Africa and then take the first 100 species. Sampling tiles from these overlapping ranges yields multiple species per tile, making this a multi-label classification task. The minimum number of tiles in which a species appears by split is 187 for train 100%, 101 for train 50%, 7 for train 5%, 32 for validation, 46 for random test, and 173 for geographic test. See the Appendix for more details.

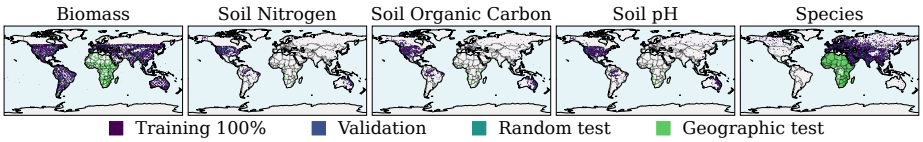

Fig. 2: Data splits in MMEarth-Bench. Each of the 5 tasks consists of a geographic test split ("Africa") and splits the rest of the world randomly into training (70%), validation (15%), and random test (15%) sets. Subsets with 50% and 5% of the training data for even lower-shot experiments can be seen with our Explorer.

4 Method

Multimodal adaptation signal. We propose a method for multimodal test-time training, TTT-MMR, in which we formulate reconstruction tasks using the *task modalities*, *i.e.* the set of modalities available at test time. As we do *not* use test labels, we treat these modalities as proxies for our task of interest. While we know that canopy height is correlated with biomass [14, 26], for example, it is generally nontrivial to preselect the best proxies for a given task. Thus, we propose a method for using all available modalities in a test-time adaptation signal. A task modality decoder is used to reconstruct all the task modalities given the encoder's embedding of the *input modalities*, *i.e.* the subset of modalities accepted by the encoder. For a given task, we first perform joint training, during which we train the task modality decoder with the mean of the modality reconstruction losses, the main task decoder with the task loss, and the encoder with the sum of these two losses. Then at test time the reconstruction losses serve as an adaptation signal for the encoder (see Fig. 3). TTT-MMR normalizes the gradients of the task modality reconstruction losses by modality and then averages them across modalities (excluding missing modalities) so that each modality contributes equally to the adaptation signal, regardless of its original scale. We freeze the task modality decoder to force the encoder to adapt.

Formalization. Each batch B contains a set of $|B|$ tiles. Each tile $t \in B$ has data for the input modalities, collectively i_t, and data for each task modality $m \in M$ given by $d_{m,t}$. The encoder is a function e with parameters θ of the input modalities, which produces input embeddings $\{e_\theta(i_t)\}_{t \in B}$. The task modality decoder is a function h with parameters α of the input embeddings, which produces modality reconstructions $\{\hat{d}_{m,t} = h_\alpha(e_\theta(i_t))_{m,t}\}_{m \in M, t \in B}$. Let $R_{m,t}(d_{m,t}, \hat{d}_{m,t})$ be the reconstruction loss for tile t's modality m. For each modality, the reconstruction loss is averaged across all tiles in the batch: $R_m = \frac{1}{|B|} \sum_{t \in B} R_{m,t}(d_{m,t}, \hat{d}_{m,t})$. We take the gradient of each per-modality reconstruction loss with respect to the encoder parameters: $\frac{\partial R_m}{\partial \theta} = \frac{\partial R_m}{\partial h_\alpha} \frac{\partial h_\alpha}{\partial e_\theta} \frac{\partial e_\theta}{\partial \theta}$. To put equal weight on each modality, we separately normalize their gradients, $\frac{\partial R_m}{\partial \theta} \to \frac{1}{\| \frac{\partial R_m}{\partial \theta} \|} \frac{\partial R_m}{\partial \theta}$, before averaging the gradients across modalities. We adapt the parameter θ_j by backpropagating this mean gradient using stochastic gradient descent with learning rate λ: $\theta_j \to \theta_j - \frac{\lambda}{|M|} \sum_{m \in M} \frac{1}{\| \frac{\partial R_m}{\partial \theta} \|} \frac{\partial R_m}{\partial \theta_j}$.

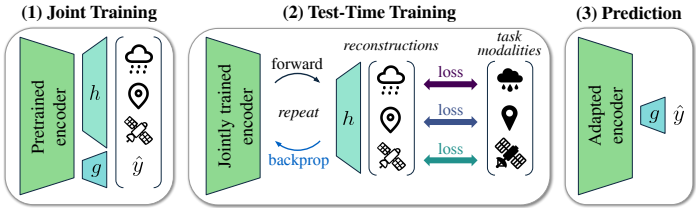

Fig. 3: TTT with multimodal reconstruction (TTT-MMR). (1) A (pretrained) *encoder* is jointly trained with the *main task decoder* g and *task modality decoder* h to get a prediction $\hat{y}$ for the main task and a reconstruction $\hat{d}_m$ for every task modality $m \in M$. Then for each batch at test time, (2) the modality reconstruction losses $\{R_m\}_{m \in M}$ are used as an adaptation signal to update the encoder iteratively, and (3) the adapted encoder is used to yield improved predictions for the main task.

In an iterative optimization process, we can use the updated encoder to generate new embeddings and reconstructions to compute new reconstruction losses and gradients, which can then update the encoder again. After TTT, the adapted encoder $e_{\theta'}$ produces embeddings that the task decoder g uses to generate a final prediction $\hat{y} = g(e_{\theta'}(i))$, after which the encoder is reset to its post-JT state for the next batch. We perform $I_{\max}$ iterations on every batch in the validation set and then use the mean of the batches' best iteration number during testing. In our experiments, we use $|B| = 8$, $\lambda = 10^{-2}$, and $I_{\max} = 5$.

Batching. We compare TTT-MMR, in which batches are random, non-overlapping samples of the test data, with TTT-MMR-Geo, which batches the test data based on geographic proximity as a proxy for sample similarity. The latter allows more specialization to the geographic domain defined by the batch, at the cost of some regularization that comes from averaging over more geographically diverse tiles. We generate the geographic batches through recursive spatial partitioning, mimicking a k-d tree [5]. We recursively subdivide the region containing the split's tiles until only one subregion has more than $|B|$ tiles. This results in non-overlapping batches that are contiguous geographic regions.

5 Experiments

We address the following research questions: (1) *How well can pretrained models transfer to downstream tasks with limited reference data?* (2) *Can pretrained models generalize geographically?* (3) *When do pretrained multimodal models benefit from using multiple input modalities at test time?* Lastly, we show that multimodal TTT improves models with limited and geographically biased data.

5.1 Setup

Benchmarking. We benchmark 8 pretrained models: 3 RGB-only, 2 S2-only, and 3 multimodal models (see Tab. 4). For comparison, we also evaluate a randomly initialized ConvNeXtV2A model with RGB input and another variant

Table 4: Pretrained models benchmarked.

Method	Architecture	Pretraining dataset	Input modalities
Scale-MAE [41]	ViT-L	FMoW-RGB [9]	RGB
DINOv3 Web [46]	ViT-L/16 distilled	LVD-1689M [46]	RGB
DINOv3 Sat [46]	ViT-L/16 distilled	SAT-493M [46]	RGB
SatlasNet [3]	Swin-v2-B	SatlasPretrain [3]	S2
MPMAE [35]	ConvNeXt V2-A	MMEarth [35]	S2
TerraMind [23]	TerraMindv1-B	TerraMesh [6]	S2, S1, DEM, RGB
Copernicus-FM [53]	ViT-B	Copernicus-Pretrain [53]	S2, S1, DEM, Geolocation, Date
Galileo [51]	ViT-B	Galileo dataset [51]	S2, S1, NDVI, Temperature, Precipitation, DEM, Dynamic World, Geolocation, Month

(ConvNeXtV2A-MM) with all modalities as input. We implement all models with a linear task decoder. For the tile-level tasks, the task decoder consists of global average pooling followed by layer norm and a fully connected layer, and for the pixel-level task, it bilinearly upsamples the embeddings to the tile size and then applies a convolutional layer with a 1×1 kernel. We finetune by training all model parameters. Linear probing results are provided in the Appendix.

Multimodal models. The TerraMind encoder contains a series of blocks in which modality-specific tokens are conditioned on one another through the attention mechanism. For modality fusion, we use the default implementation, which averages the patch embeddings from the final encoder block across modalities. In contrast, cross-modal Copernicus-FM does not have per-modality encoding. We implement it as a Siamese network, sharing parameters across modalities. We separately compute embeddings for S2, S1, and DEM and then average them. Galileo linearly projects the modalities separately into tokens, adds modality-specific embeddings to them, concatenates all the tokens into a single sequence, and then performs self-attention in shared transformer blocks. For ConvNeXtV2A-MM we one-hot-encode the categorical modalities, add spatial dimensions to the tile-level modalities, stack all the pixel- and tile-level bands in a single tensor, and set the number of input channels to 926.

JT and TTT. JT and TTT use a linear task modality decoder that bilinearly upsamples the input embeddings to the tile size and then passes them through a 2D convolution with a 1×1 kernel that reconstructs every band in every task modality at the tile resolution. We group the channels by modality and perform global average pooling on the bands of the tile-level modalities, yielding the modality reconstructions. For the reconstruction losses we use cross-entropy loss and MSE loss for the categorical- and continuous-valued modalities, respectively.

Implementation. We run our experiments on an NVIDIA H200 140GB GPU. The regression tasks use MSE loss and species uses BCE with logits loss. Our hyperparameter settings are specified in the Appendix. In all results, "performance" refers to R^2 for the regression tasks and mean average precision (mAP) for species. Following training, we select the checkpoint with the highest performance on the validation set and report its results on both test sets. The input data is normalized according to the pretrained model's normalization method and statistics or center-normalized using the mean and STD of the bands in the

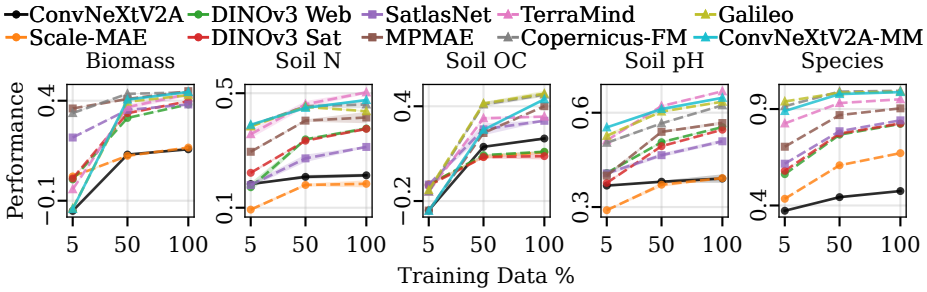

Fig. 4: Low-shot in-distribution performance. Finetuning on training subsets. Symbology: ●=RGB, ■=S2, ▲=multimodal, solid=random init., dashed=pretrained.

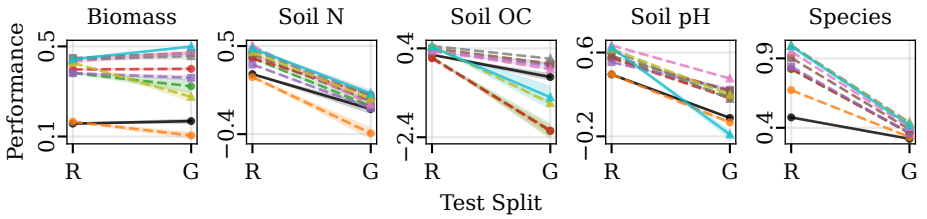

Fig. 5: Geographic generalization. Performance comparison on random (R) *vs.* geographic (G) test splits using all training data.

training set for the randomly initialized models and the task modalities. Shaded regions in plots reflect 1 standard error across three distinct random seeds.

5.2 Results

Finetuning with limited reference data. We finetune all 8 pretrained models and the randomly initialized ConvNeXtV2A models on three subsets of the training data: 5%, 50%, and 100%. We evaluate the models on the random, in-distribution test split to determine how effectively they transfer to the five tasks in MMEarth-Bench, with results shown in Fig. 4. Overall, the models that take multimodal input data outperform the unimodal models and especially the RGB-only models. MPMAE, an S2-only model, shows competitive or superior performance to the multimodal models on the biomass and soil OC tasks especially. Surprisingly, the randomly initialized ConvNeXtV2A-MM model generally performs on par with the pretrained multimodal models, ranking 1st when averaging rankings across tasks for 100% training data (see Appendix). As expected, its performance relative to the pretrained multimodal models worsens with less training data, but it still ranks 4th overall with 5% training data, only beaten by the other multimodal models. The RGB-only ConvNeXtV2A model ranks poorly overall, but it performs similarly to or better than at least one of the RGB-pretrained models on 4 out of 5 tasks. Biomass and soil OC appear to be most sensitive to limited training data, with none of the models achieving a

positive R^2 with only 5% of the training data on the soil OC task, which existing work in soil modeling has found to be challenging [15]. Comparing DINOv3 Web and Sat shows that the domain-specific pretraining of DINOv3 Sat had little benefit on our tasks. Copernicus-FM originally used multimodal inputs during linear probing and unimodal inputs during finetuning in a cross-modal evaluation [53]. We provide a *multimodal* finetuning evaluation of Copernicus-FM.

Geographic generalization challenge. Next, we evaluate all the models on the geographic test split after finetuning on all the training data to determine how effectively they generalize to Africa, for which they have not seen labels. As shown in Fig. 5, for all tasks except biomass the models perform significantly worse on the geographic test split than on the random one, despite being pretrained globally. Overall, the multispectral models perform best, though for soil N and soil OC, even the multimodal models struggle on the geographic split, suggesting that these properties are especially challenging to predict in new regions. The randomly initialized ConvNeXtV2A-MM ranks 4th overall, while the multimodal pretrained Galileo model ranks 5th. ConvNeXtV2A-MM's lower ranking on the geographic than the random test split can be explained by i) it has not seen any pretraining data from Africa and ii) being trained on all the modalities may increase the risk of overfitting to the training distribution. This means that the representations learned for Africa during pretraining can be beneficial when labeled data is unavailable, especially for soil OC and pH. The DINOv3 models have likely also seen unlabeled data from Africa during pretraining, but they are still outperformed by ConvNeXtV2A-MM on 4 of the 5 tasks on the geographic split. ConvNeXtV2A-MM's strong performance on biomass and soil N suggests that the multimodal inputs are helping to resolve ambiguities that pretraining is not able to achieve. It is also possible that finetuning on the world without Africa erases some of the information the pretrained models had encoded about the region, hindering geographic generalization. All the soil tasks and species occurrence offer an opportunity for globally pretrained models to improve their methodology to better facilitate geographic generalization on downstream tasks.

Unimodal *vs*. multimodal input data. We compare multimodal and S2-only variants of TerraMind, Copernicus-FM, and Galileo to evaluate the effect of multimodal input data on their performance. Fig. 6 shows that the models can use the additional modalities to produce better task-specific representations on both test splits. While the multimodal version usually outperforms the unimodal one (less pronounced for TerraMind), there are exceptions in the geographic split of the soil tasks. While Fig. 5 showed that multimodal pretraining tends to lead to better geographic generalization than unimodal pretraining, Fig. 6 suggests that *finetuning* the same multimodally pretrained model with multi- rather than unimodal inputs can harm generalization. The additional modalities could lead to overfitting to the non-Africa training domain, which is harmful if the modalities undergo domain shifts between non-Africa and Africa. This effect appears more strongly for TerraMind and Galileo than Copernicus-FM, which could be because our Siamese Copernicus-FM simply averages embeddings across modalities, but the fusion is learned in the other two. This behavior has been

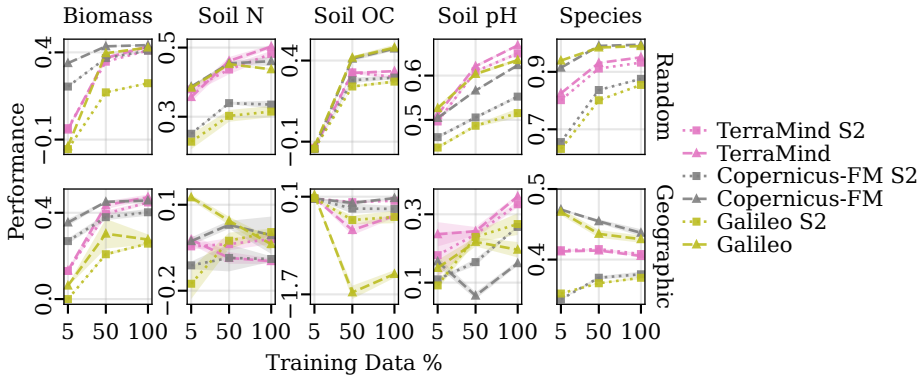

Fig. 6: Unimodal *vs*. multimodal input data. Finetuning performance of S2-only (dotted-square) *vs*. multimodal (dashed-triangle) variants.

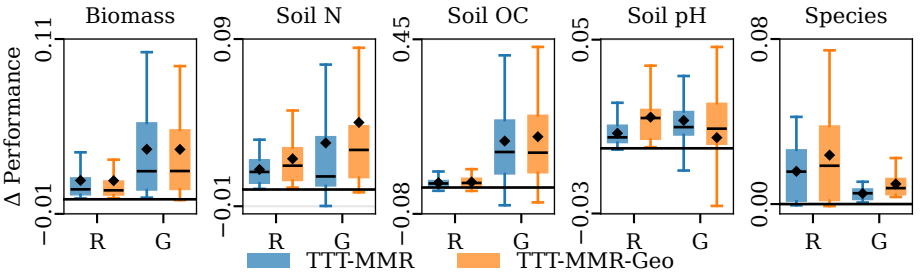

Fig. 7: Multimodal test-time training improvement per task. Improvement of TTT-MMR (random batching) and TTT-MMR-Geo (geographic batching) over joint training. Boxplots show the distribution over all models and seeds. ♦=mean, −=median, whiskers are based on the 1.5 IQR value. Δ reflects absolute change.

observed in prior work [40]. Furthermore, TerraMind and Galileo have modality-specific encoders and thus more capacity to fit the training data.

Multimodal test-time training. After performing JT, we apply TTT-MMR or TTT-MMR-Geo to the models to improve performance. Fig. 7 shows that both of our proposed approaches improve performance over JT on all tasks. For biomass, soil OC, and soil pH, both variants perform similarly, with TTT-MMR-Geo displaying better performance on soil N and species. One-sided (greater) Wilcoxon tests with Holm-Bonferroni correction demonstrate statistical significance of the performance improvement after TTT for each boxplot with $p < 0.05$. In addition, both methods improve performance on both test splits, with larger gains on the geographic split for biomass and soil OC.

The method rankings per model averaged over tasks shown in Tab. 5 demonstrate that both our TTT methods outperform JT for all models. TTT-MMR-Geo is the best method for all models except SatlasNet, Galileo, and ConvNeXtV2A-MM, indicating that these may benefit from the additional regular-

Table 5: Multimodal TTT improvement per model. Average ranks of JT (joint training baseline), TTT-MMR (random batching), and TTT-MMR-Geo (geographic batching). Ranks are mean ± standard error averaged over tasks and seeds.

Test split	Method	ConvNeXt V2A	Scale-MAE	DINOv3 Web	DINOv3 Sat	Satlas Net	MPMAE	Terra Mind	Copernicus -FM	Galileo	ConvNeXt V2A-MM
Random	JT	2.9 ± 0.1	3.0 ± 0.0	3.0 ± 0.0	3.0 ± 0.0	2.9 ± 0.1	3.0 ± 0.0	3.0 ± 0.0	2.8 ± 0.1	2.2 ± 0.2	2.5 ± 0.2
	TTT-MMR	2.1 ± 0.1	1.8 ± 0.1	1.8 ± 0.1	1.7 ± 0.1	1.8 ± 0.1	1.6 ± 0.1	**1.5 ± 0.1**	1.7 ± 0.2	**1.7 ± 0.1**	**1.7 ± 0.2**
	TTT-MMR-Geo	**1.1 ± 0.1**	**1.2 ± 0.1**	**1.2 ± 0.1**	**1.3 ± 0.1**	**1.3 ± 0.1**	**1.4 ± 0.1**	**1.5 ± 0.1**	**1.5 ± 0.1**	2.1 ± 0.2	1.8 ± 0.2
Geographic	JT	3.0 ± 0.0	2.8 ± 0.1	3.0 ± 0.0	3.0 ± 0.0	2.7 ± 0.2	2.7 ± 0.2	2.9 ± 0.1	2.8 ± 0.1	2.9 ± 0.1	2.5 ± 0.2
	TTT-MMR	**1.5 ± 0.1**	1.9 ± 0.2	1.7 ± 0.1	**1.5 ± 0.1**	**1.6 ± 0.2**	2.0 ± 0.1	1.7 ± 0.1	1.8 ± 0.1	1.6 ± 0.1	1.9 ± 0.1
	TTT-MMR-Geo	**1.5 ± 0.1**	**1.3 ± 0.1**	**1.3 ± 0.1**	**1.5 ± 0.1**	1.7 ± 0.2	**1.3 ± 0.2**	**1.3 ± 0.2**	**1.4 ± 0.2**	**1.5 ± 0.2**	**1.7 ± 0.2**

ization that comes from TTT-MMR's random batching. The RGB-only models tend to undergo larger performance improvements from TTT than the multispectral and multimodal models (shown in the Appendix). These are the models that otherwise have the least 'information' about a given tile, but both TTT-MMR variants allow them to make use of multimodal data. The pretrained S2 and multimodal models still get additional modalities through TTT-MMR but also get more spectral information as input. Although MPMAE was pretrained by reconstructing these same modalities, we do not observe any special behavior in its performance gains from TTT. Overall, all benchmarked models benefit from our methods, despite having diverse architectures, pretraining strategies, and input modalities. This demonstrates that our TTT methods are model- and task-agnostic. Combining all 12 task modality reconstruction losses helps prevent overfitting to any single reconstruction task, producing a more robust adaptation signal that is more likely to align with the actual task of interest. Moreover, our per-modality gradient normalization improves stability during TTT.

Long-tail performance improves especially with TTT-MMR-Geo. Stratifying the same TTT results by sample frequency in Fig. 8 reveals that geographic batching plays a crucial role in improving performance on the long-tail of the target distributions. For the regression tasks, as the sample frequency decreases, JT and TTT-MMR increasingly underestimate the target value, whereas TTT-MMR-Geo is robust to sample rarity. This can be explained by the greater specialization provided by geographic batching. Tiles located closer together are more likely to be more similar, so the tiles' gradients used for adapting the encoder are more likely to point in similar directions. For random batching, averaging over the tiles' gradients means any rare tile's effect gets diluted. For the species classification task, the three methods' robustness to sample frequency is more similar, but for the random split we observe greater performance gains from TTT-MMR-Geo for rarer species than more common ones. This may be due to the random test set being more geographically dispersed, so a random batch is likely more diverse than a random batch within Africa.

6 Conclusion

Models pretrained on large, unlabeled Earth observation datasets aim to generalize to new tasks and geographic domains with limited data. To evaluate

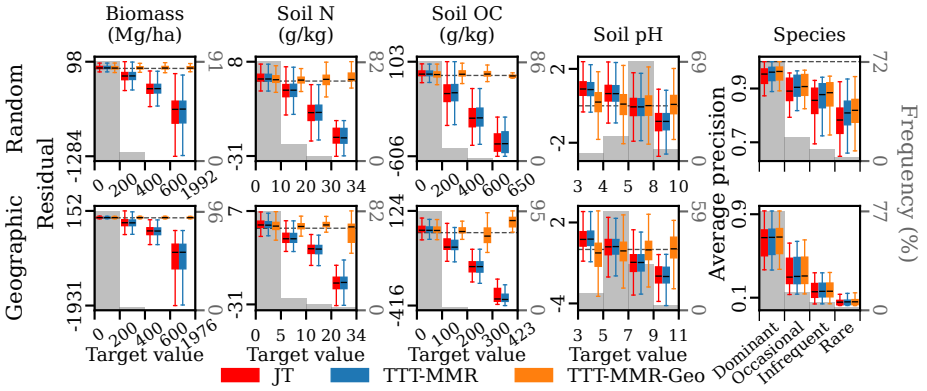

Fig. 8: Long-tail analysis of multimodal test-time training. Residuals and average precision (per species) stratified by sample frequency in each test split. Geographic batching (TTT-MMR-Geo) substantially improves upon joint training (JT) and random batching (TTT-MMR) for rare samples. Boxplots show the distribution over all models with one seed using the median and whiskers based on the 1.5 IQR value.

their progress, we present MMEarth-Bench, a new multimodal benchmark dataset that contains 12 aligned EO modalities and task data for biomass, soil nitrogen, soil organic carbon, soil pH, and species occurrence. We benchmark 8 pretrained (RGB, multispectral, and multimodal) models on these tasks and find that all models exhibit a geographic generalization gap on 4 out of 5 tasks. Self-supervised multimodal pretraining improves downstream performance in low-shot regimes and under geographic shifts, but with sufficient labels a simple multimodal model trained from scratch is competitive. To enable *any* (pretrained) model to employ all available modalities when making predictions, we propose self-supervised multimodal *test-time* training. Our proposed TTT-MMR method uses all modalities as reconstruction tasks at test time to provide a self-supervised adaptation signal. We show that TTT-MMR can improve performance on all models, tasks, and splits over a joint training baseline, with geographic batching displaying unique robustness to rare samples. Thus, we have formulated an efficient, model-agnostic method for improving adaptation performance at test time.

Acknowledgments

We appreciate the open data policies of the Copernicus program and its partners ESA and ECMWF. We thank Harvard's FAS Research Computing cluster and also Google Earth Engine for free access to the data. This work was supported by the Pioneer Centre for AI (DNRF grant no. P1), research grant Global Wetland Center (grant no. NNF23OC0081089) from the Novo Nordisk Foundation, and the European Union project ELIAS (grant agreement no. 101120237). LG was supported by the National Science Foundation Graduate Research Fellowship (grant no. DGE 2140743) and VILLUM FONDEN (grant no. VIL70006).

References

1. Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities. In: CVPR (2025). https://doi.org/10.1109/CVPR52734.2025.01819
2. Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. ICCV (2021)
3. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In: ICCV. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01538
4. Batjes, N.H., Calisto, L., de Sousa, L.M.: Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). Earth System Science Data (2024). https://doi.org/10.5194/essd-16-4735-2024
5. Bentley, J.L.: Multidimensional binary search trees used for associative searching. Commun. ACM (1975), https://api.semanticscholar.org/CorpusID:13091446
6. Blumenstiel, B., Fraccaro, P., Marsocci, V., Jakubik, J., Maurogiovanni, S., Czerkawski, M., Sedona, R., Cavallaro, G., Brunschwiler, T., Bernabe-Moreno, J., Longépé, N.: TerraMesh: A Planetary Mosaic of Multimodal Earth Observation Data (2025), https://arxiv.org/abs/2504.11172
7. Bountos, N.I., Ouaknine, A., Papoutsis, I., Rolnick, D.: FoMo: multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. In: AAAI. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025). https://doi.org/10.1609/aaai.v39i27.35002
8. Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., et al.: Dynamic world, near real-time global 10 m land use land cover mapping. Scientific data (2022)
9. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional Map of the World. In: CVPR (2018). https://doi.org/10.1109/CVPR.2018.00646
10. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S.: SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) NeurIPS (2022), https://openreview.net/forum?id=WBhqzpF6KYH
11. iNaturalist contributors: iNaturalist Research-grade Observations. Occurrence dataset https://doi.org/10.15468/ab3s5x accessed via GBIF.org on 2025-11-10 (2025)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009). https://doi.org/10.1109/CVPR.2009.5206848
13. Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., et al.: An ecoregion-based approach to protecting half the terrestrial realm. BioScience (2017)
14. Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., et al.: The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. Science of remote sensing (2020)
15. Feeney, C., Cosby, B., et al., D.R.: Multiple soil map comparison highlights challenges for predicting topsoil organic carbon concentration at national scale. Sci Rep 12 **1379** (2022), https://doi.org/10.1038/s41598-022-05476-5

16. Fibaek, C., Camilleri, L., Luyts, A., Dionelis, N., Saux, B.L.: PhilEO Bench: Evaluating Geo-Spatial Foundation Models. In: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium (2024). https://doi.org/10.1109/IGARSS53475.2024.10642780
17. Gandelsman, Y., Sun, Y., Chen, X., Efros, A.A.: Test-Time Training with Masked Autoencoders. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) NeurIPS (2022), https://openreview.net/forum?id=SHMi1b7sjXk
18. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment (2017)
19. Goyal, S., Sun, M., Raghunanthan, A., Kolter, Z.: Test-time adaptation via conjugate pseudo-labels. NeurIPS (2022)
20. Hendrycks, D., Lee, K., Mazeika, M.: Using Pre-Training Can Improve Model Robustness and Uncertainty. In: ICML (2019)
21. Hong, D., Li, C., Li, X., Camps-Valls, G., Chanussot, J.: Foundation models in remote sensing: Evolving from unimodality to multimodality. IEEE Geoscience and Remote Sensing Magazine (2026)
22. IUCN: The IUCN Red List of Threatened Species. Version 2025-1 **Downloaded on June 13, 2025** (2025)
23. Jakubik, J., Yang, F., Blumenstiel, B., Scheurer, E., Sedona, R., Maurogiovanni, S., Bosmans, J., Dionelis, N., Marsocci, V., Kopp, N., et al.: TerraMind: Large-Scale Generative Multimodality for Earth Observation. ICCV (2025)
24. Jeong, W., Cho, J., Yoon, Y., Yoon, K.J.: Synchronizing Task Behavior: Aligning Multiple Tasks during Test-Time Training. In: ICCV (2025)
25. Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E.D., Kerner, H., Lütjens, B., Irvin, J.A., Dao, D., Alemohammad, H., Drouin, A., Gunturkun, M., Huang, G., Vazquez, D., Newman, D., Bengio, Y., Ermon, S., Zhu, X.X.: GEO-Bench: Toward Foundation Models for Earth Monitoring. NIPS (2023)
26. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the Earth. Nature Ecology & Evolution (2023)
27. Li, W., Chen, K., Chen, H., Shi, Z.: Geographical knowledge-driven representation learning for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2021)
28. Liu, X., Guo, Z., Li, S., Xing, F., You, J., Kuo, C.C.J., El Fakhri, G., Woo, J.: Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In: ICCV (2021). https://doi.org/10.1109/ICCV48922.2021.01020
29. Liu, X., Yoo, C., Xing, F., Oh, H., Fakhri, G., Kang, J.W., Woo, J.: Deep unsupervised domain adaptation: A review of recent advances and perspectives. APSIPA Transactions on Signal and Information Processing (2022). https://doi.org/10.1561/116.00000192
30. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: ICCV (2021)
31. 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)
32. Ming, D., Tan, S., Shailesh, Liu, B., Raj, A., Ang, Q.X., Dai, W., Duhan, T., Chiun, J., Cao, Y., Shkurti, F., Sartoretti, G.: Search-tta: A multimodal test-time adaptation framework for visual search in the wild. In: Conference on Robot Learning. PMLR (2025)

33. Mizrahi, D., Bachmann, R., Kar, O.F., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4M: Massively Multimodal Masked Modeling. In: NeurIPS (2023)
34. NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team: Digital Elevation Model V003. https://doi.org/10.5067/ASTER/ASTGTM.003 (2018)
35. Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., Lang, N.: MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In: ECCV. Springer (2024)
36. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards Stable Test-Time Adaptation in Dynamic Wild World. In: ICLR (2023)
37. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
38. Persello, C., Wegner, J.D., Hänsch, R., Tuia, D., Ghamisi, P., Koeva, M., Camps-Valls, G.: Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. IEEE Geoscience and Remote Sensing Magazine (2022)
39. Picek, L., Botella, C., Servajean, M., Leblanc, C., Palard, R., Larcher, T., Deneu, B., Marcos, D., Bonnet, P., Joly, A.: GeoPlant: Spatial Plant Species Prediction Dataset. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) NIPS. vol. 37. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-4023
40. Rao, A., Rolf, E.: Using Multiple Input Modalities can Improve Data-Efficiency and O.O.D. Generalization for ML with Satellite Imagery. In: TerraBytes - ICML 2025 workshop (2025), https://www.arxiv.org/abs/2507.13385
41. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In: ICCV. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00378
42. Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., Jacobs, N.: Taxabind: A unified embedding space for ecological applications. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025). https://doi.org/10.1109/WACV61041.2025.00179
43. Scheibenreif, L., Mommert, M., Borth, D.: Parameter efficient self-supervised geospatial domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
44. Shin, I., Tsai, Y.H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I.S., Yoon, K.J.: MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In: CVPR (2022). https://doi.org/10.1109/CVPR52688.2022.01642
45. Sialelli, G., Peters, T., Wegner, J.D., Schindler, K.: AGBD: A Global-scale Biomass Dataset. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-G-2025** (2025). https://doi.org/10.5194/isprs-annals-x-g-2025-829-2025
46. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
47. Su, J.C., Maji, S., Hariharan, B.: When Does Self-supervision Improve Few-Shot Learning? In: ECCV. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58571-6_38

48. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision. arXiv preprint arXiv:1909.11825 (2019)
49. Sun, Y., Wang, X., Zhuang, L., Miller, J., Hardt, M., Efros, A.A.: Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In: ICML (2020)
50. Teng, M., Elmustafa, A., Akera, B., Bengio, Y., Radi, H., Larochelle, H., Rolnick, D.: SatBird: a Dataset for Bird Species Distribution Modeling using Remote Sensing and Citizen Science Data. In: NIPS (2023), https://openreview.net/forum?id=Vn5qZGxGj3
51. Tseng, G., Fuller, A., Reil, M., Herzog, H., Beukema, P., Bastani, F., Green, J.R., Shelhamer, E., Kerner, H., Rolnick, D.: Galileo: Learning Global & Local Features of Many Remote Sensing Modalities. In: ICML (2025), https://openreview.net/forum?id=gqZO3eSZRy
52. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully Test-Time Adaptation by Entropy Minimization. In: ICLR (2021), https://openreview.net/forum?id=uXl3bZLkr3c
53. Wang, Y., Xiong, Z., Liu, C., Stewart, A.J., Dujardin, T., Bountos, N.I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., Zhu, X.X.: Towards a Unified Copernicus Foundation Model for Earth Vision (2025), https://arxiv.org/abs/2503.11849
54. Wang, Z., Zhang, Y., Zhang, Z., Jiang, Z., Yu, Y., Li, L., Zhang, L.: Exploring Uncertainty-Based Self-Prompt for Test-Time Adaptation Semantic Segmentation in Remote Sensing Images. Remote Sensing **16**(7) (2024). https://doi.org/10.3390/rs16071239
55. WEF: Unlocking the potential of earth observation to address africa's critical challenges. Tech. rep., World Economic Forum (2021), https://www3.weforum.org/docs/WEF_Digital_Earth_Africa_Unlocking_the_potential_of_Earth_Observation_to_address_Africa_2021.pdf, in collaboration with Digital Earth Africa
56. Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L., Camps-Valls, G., Zhu, X.X.: Neural Plasticity-Inspired Foundation Model for Observing the Earth Crossing Modalities. arXiv preprint arXiv:2403.15356 (2024)
57. Yang, M., Li, Y., Zhang, C., Hu, P., Peng, X.: Test-time adaptation against multimodal reliability bias. In: ICLR (2024), https://openreview.net/forum?id=TPZRq4FALB
58. Yau, E.Y., Jones, E.E., Tsang, T.P., Xing, S., Corlett, R.T., Roehrdanz, P., Lohman, D.J., Lee, A.K., Hai, C.W., Chowdhury, S.e.a.: Spatial occurrence records and distributions of tropical asian butterflies. Scientific Data **12**(1004) (Jun 2025). https://doi.org/10.32942/x2c904
59. Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D.B., Ermon, S.: SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In: NeurIPS (12 2021), https://openreview.net/forum?id=5HR3vCylqD
60. Yeo, T., Kar, O.F., Sodagar, Z., Zamir, A.: Rapid Network Adaptation: Learning to Adapt Neural Networks Using Test-Time Feedback. In: ICCV (2023)
61. Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.E., Ramoino, F., Arino, O.: ESA WorldCover 10 m 2020 v100 (2021). https://doi.org/10.5281/zenodo.5571936

62. Zhao, S., Li, S.Y., Huang, S.J.: NanoAdapt: Mitigating Negative Transfer in Test Time Adaptation with Extremely Small Batch Sizes. In: Larson, K. (ed.) IJCAI. International Joint Conferences on Artificial Intelligence Organization (8 2024). https://doi.org/10.24963/ijcai.2024/616, Main Track
63. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine (2017). https://doi.org/10.1109/MGRS.2017.2762307