Towards a Unified Copernicus Foundation Model for Earth Vision

Supplementary Material

A. Copernicus-Pretrain

This section reports more detailed characteristics and statistical analyses for the Copernicus-Pretrain dataset.

A.1. Comparison to existing EO pretraining datasets

Tab. 1 shows a detailed comparison between Copernicus-Pretrain and several existing EO pretraining datasets.

Table 1. A comparison of existing EO pretraining datasets.

Dataset	Modality	Resolution	# Time stamps	# patches	# pixels
fMoW [8]	RGB, MS	0.3–10 m	3	2M	50B
SEN12MS [23]	SAR, MS	10 m	1	540K	35B
SeCo [18]	MS	10 m	5	1M	70B
SSL4EO-S12 [26]	SAR, MS	10 m	4	3M	140B
SSL4EO-L [24]	MS	30 m	4	5M	348B
SatlasPretrain [2]	SAR, MS, RGB	0.5–10 m	$\sim$10	$>$10M	17T
MMEarth [19]	SAR, MS, height, landcover, etc.	10–15 m	1	6M	120B
SpectralEarth [6]	HS	30 m	1–23	540K	10B
Major TOM [12]	SAR, MS	10 m	1	8M	6.8T
Copernicus-Pretrain	SAR, MS, S3, DEM, S5P	10 m–1 km	1–12	19M	920B

A.2. Extended statistics

All-modality-aligned subset The Copernicus-Pretrain dataset contains 310K grids with at least one modality, of which 220K have all eight modalities. Tab. 2 shows the detailed characteristics of the 220K subset, and Fig. 1 presents its global distribution. We refer to the full dataset (grids with at least one modality) as “union”, and the all-modality-aligned subset (grids with all modalities) as “joint”.

Table 2. Copernicus-Pretrain dataset characteristics (joint 220K subset).

	image size	# grid cells	# patches	# timestamps	# total images
Sentinel-1 GRD	$\sim$264x264	219,543	996,978	$\sim$4	3,948,217
Sentinel-2 TOA	$\sim$264x264	219,543	996,978	$\sim$4	3,948,217
Sentinel-3 OLCI	$\sim$96x96	219,543	219,543	$\sim$8	1,720,881
Sentinel-5P CO	$\sim$28x28	219,543	219,543	1–12	1,548,349
Sentinel-5P NO2	$\sim$28x28	219,543	219,543	1–12	1,394,800
Sentinel-5P SO2	$\sim$28x28	219,543	219,543	1–12	1,188,864
Sentinel-5P O3	$\sim$28x28	219,543	219,543	1–12	1,750,542
Copernicus DEM	$\sim$960x960	219,543	219,543	1	219,543
Copernicus-Pretrain	-	219,543	3,311,214	-	15,720,353

Statistics of local patches Fig. 2 shows the histograms of the number of local patches across grids for S1/2 in the full datasets (union), and Fig. 3 shows the histograms for the joint subset.

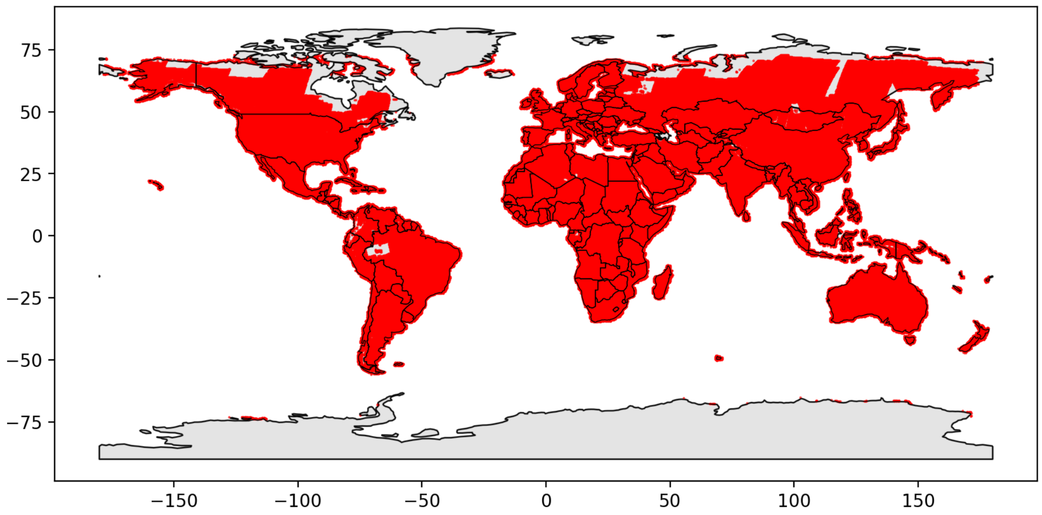


Figure 1. Global distribution of the joint subset of the Copernicus-Pretrain dataset.

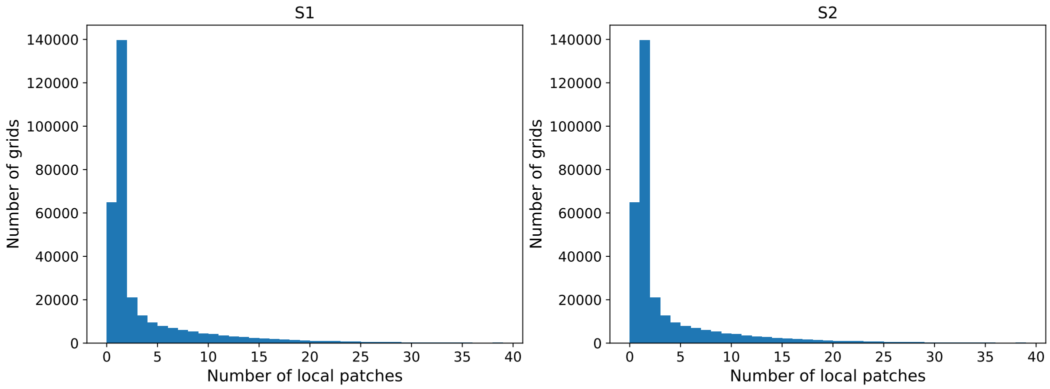


Figure 2. Histogram of local patch numbers for S1 and S2 (union).

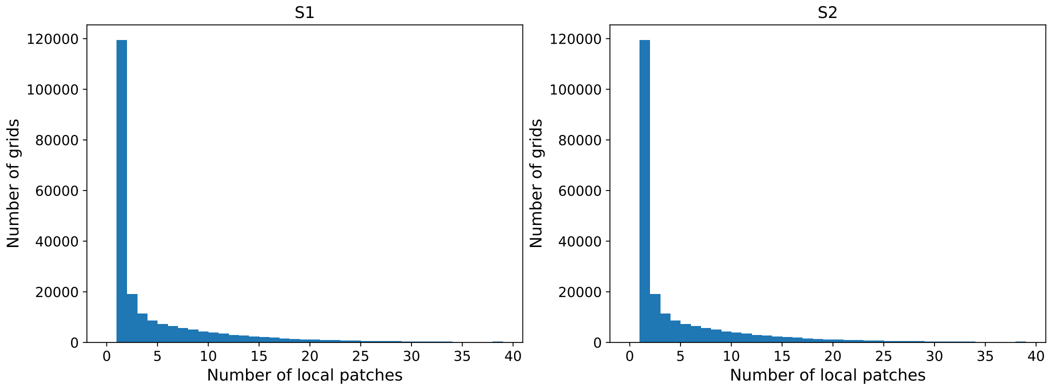


Figure 3. Histogram of local patch numbers for S1 and S2 (joint).

Statistics of time series. Fig. 4 presents the histograms of the time series lengths for S1 and S2 in the full dataset, while Fig. 5 shows the corresponding histograms in the joint subset. Similarly, Fig. 6 (left and right) presents S3 in the full dataset and joint subset, and Figs. 7 and 8 present S5P in the full dataset and joint subset.

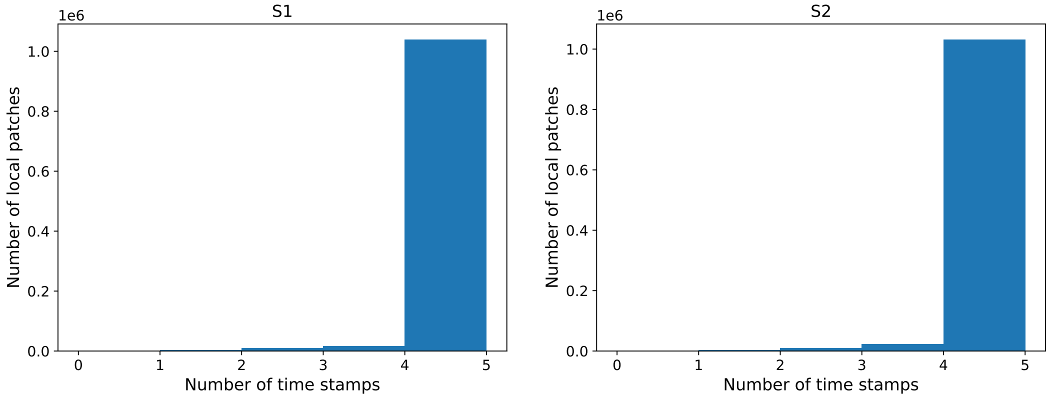


Figure 4. Histogram of time series lengths for S1 and S2 (union).

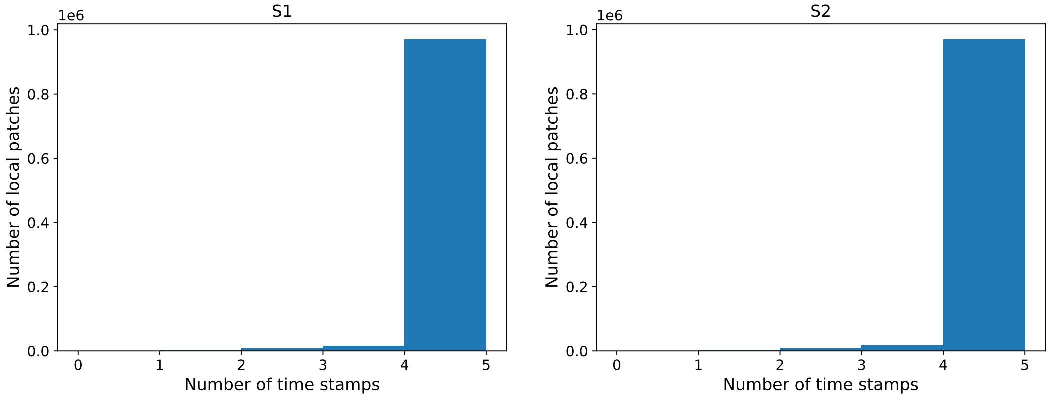


Figure 5. Histogram of time series lengths for S1 and S2 (joint).

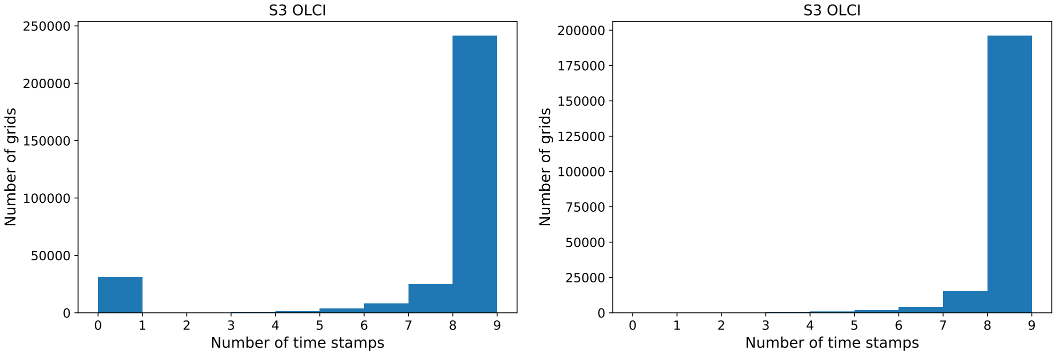


Figure 6. Histogram of time series lengths for S3 (left: union; right: joint).

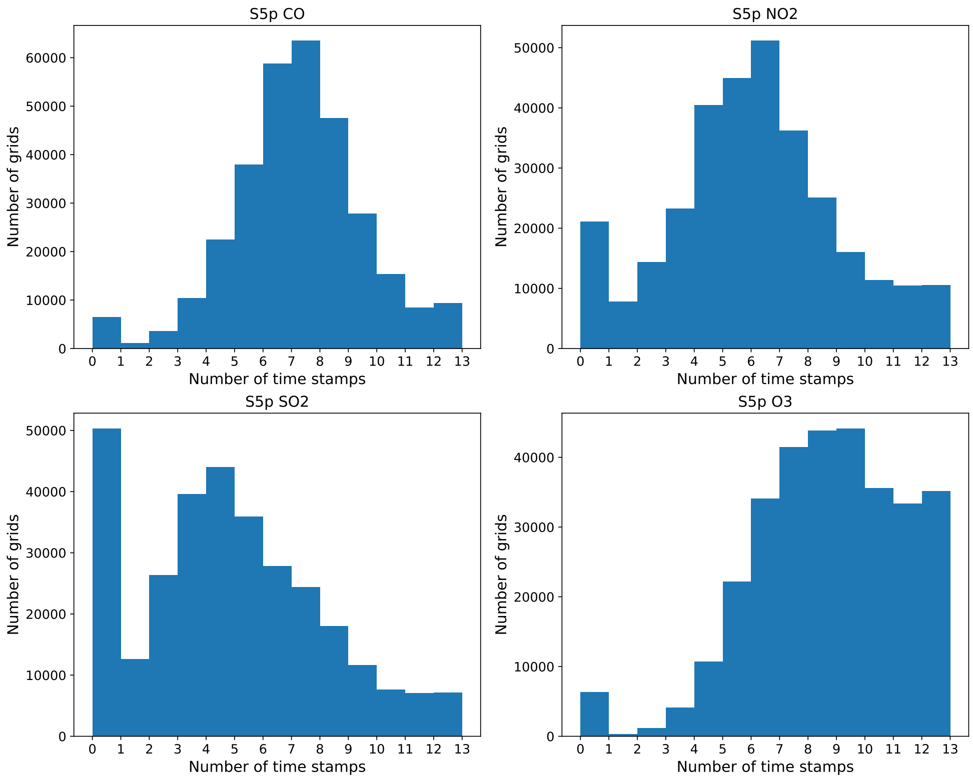


Figure 7. Histogram of time series lengths for S5P (union).

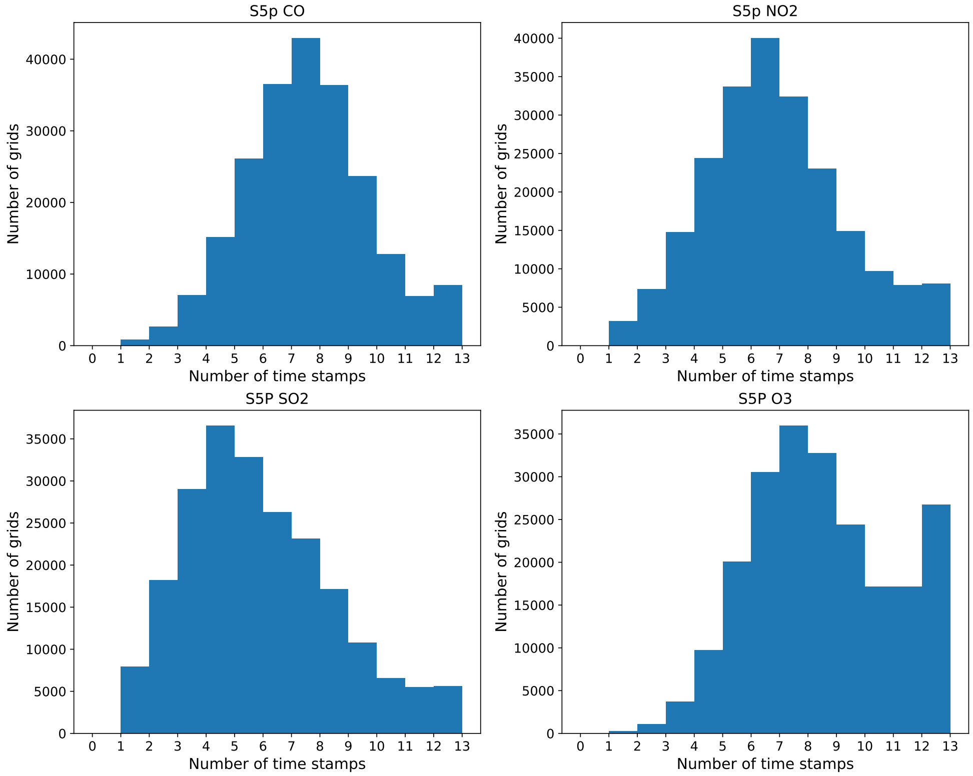


Figure 8. Histogram of time series lengths for S5P (joint).

B. Copernicus-FM

This section reports more implementation details, analyses, visualizations, and ablation studies for the Copernicus-FM foundation model. Unless explicitly noticed, for most ablation experiments, we pretrain a ViT-Small on a 10K-grid subset of Copernicus-Pretrain for 100 epochs with continual distillation only from DINOv2 [21] for efficiency.

B.1. Subtractive ablation

The incremental ablation in the main paper demonstrates the design evolution, but does not isolate individual contributions. To complement, we conduct a subtractive ablation in Table 3, showing the benefits of each component regardless of order.

Table 3. Subtractive ablation. w/o means without.

	EU-S1	EU-S2	EU-RGB	LC-S3	O3-S5P (↓)
Copernicus-FM	81.0	89.5	78.9	90.7	811.6
w/o var hypernet	78.9	87.9	78.6	90.5	857.6
w/o metadata	56.9	88.3	70.1	86.9	1556.3
w/o distill	77.9	88.9	78.5	90.7	839.3

B.2. Spectral hypernetwork

We use a unified Fourier encoding [3] to encode wavelengths and bandwidths for all spectral channels, which are added together and serve as input to the spectral hypernetwork to generate patch embedding weights.

Wavelength and bandwidth details Tab. 4 lists the detailed wavelength and bandwidth values for each spectral sensor in the Copernicus-Pretrain dataset used during our Copernicus-FM pretraining.

Sensor	Wavelengths (nm)	Bandwidths (nm)
S1 GRD	5e7, 5e7	1e9, 1e9
S2 TOA	440, 490, 560, 665, 705, 740, 783, 842, 860, 940, 1370, 1610, 2190	20, 65, 35, 30, 15, 15, 20, 115, 20, 20, 30, 90, 180
S3 OLCI	400, 412.5, 442.5, 490, 510, 560, 620, 665, 673.75, 681.25, 708.75, 753.75, 761.25, 764.375, 767.5, 778.75, 865, 885, 900, 940, 1020	15, 10, 10, 10, 10, 10, 10, 10, 7.5, 7.5, 10, 7.5, 7.5, 3.75, 2.5, 15, 20, 10, 10, 20, 40

Table 4. Wavelengths and bandwidths for different spectral sensors in Copernicus-FM pretraining.

Fourier encoding visualization Fig. 9 illustrates the Fourier encoded wavelengths and bandwidths (with 128 feature dimensions) for 13 S2 bands and 1 S1 band.

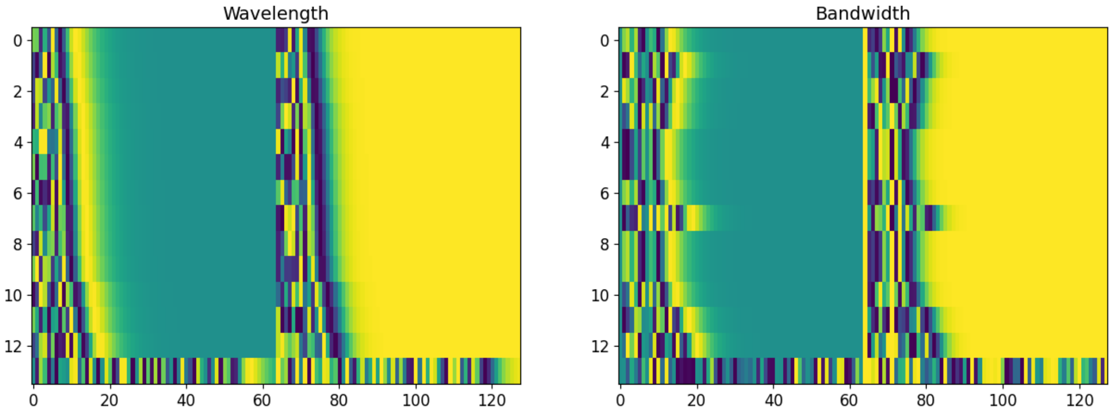


Figure 9. Fourier encoding visualization for wavelengths and bandwidths of S2 and S1.

B.3. Variable hypernetwork

We use a large language model with general cross-domain knowledge to encode variable names for non-spectral modalities. The resulting variable encodings serve as input to the variable hypernetwork to generate patch embedding weights.

Language encoding visualization Fig. 10 presents a t-SNE plot of Llama-3.2-encoded variable names. We compare the variable names in our pretraining dataset with other out-of-domain concepts. The figure indicates that the language model does have meaningful knowledge of these different names — S5P variables are gathered together, EO modalities are far away from other domains like games or mountains, and concepts within a subdomain are further well clustered.

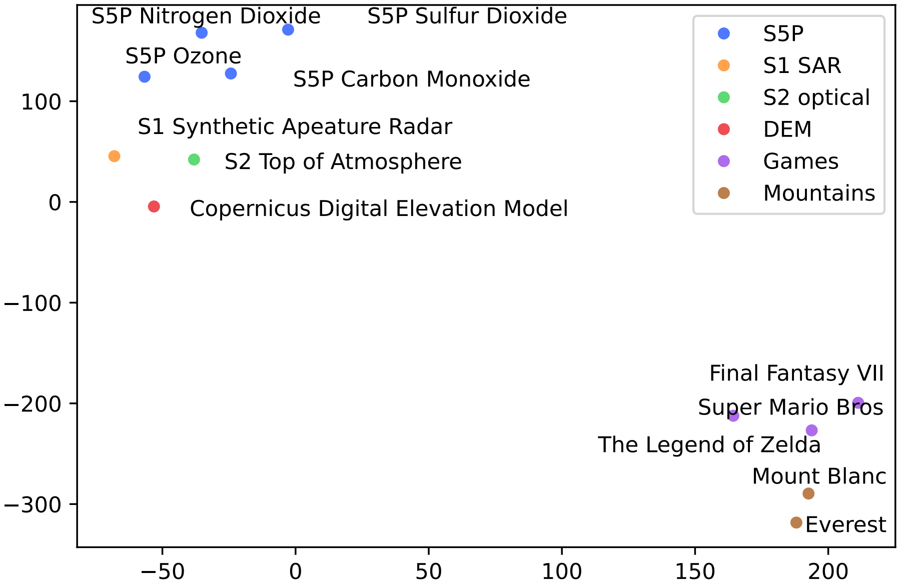


Figure 10. t-SNE visualization of the language encodings of different variable names.

Ablation on different variable encoding options Language encoding maximizes the flexibility to process any variable names, and also maintains semantic relationships between variables. Besides that, random hashing and spectral-sensitivity-guided spectroscopy are two other options to encode different variables. However, compared to language encodings, random hashing loses the flexibility and semantic relationship, while spectroscopy has semantics but lacks flexibility. Quantitatively, we conduct a comparison study in Table 5, suggesting the superior and stable performance of language encoding.

Table 5. Variable name encoding ablation.

	EU-S1	EU-S2	EU-RGB	LC-S3	O3-S5P (↓)
LLM (LLaMa-3.2-1B)	**81.0**	**89.5**	78.9	**90.7**	**811.6**
Random hash	77.7	89.5	78.6	89.9	818.6
Spectroscopy	80.4	89.5	**79.5**	89.9	815.7

B.4. Metadata integration

We use a unified Fourier encoding [3] to integrate metadata as encoding vectors added to the positional encodings.

Fourier encoding visualization Figs. 11 and 12 illustrate the Fourier encoded metadata (location, area, and time) for a few representative example values as below:

- location (lon + 180°): $0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°, 360°, 360°$;

- location (lat + 90°): $0°, 45°, 90°, 135°, 180°, 0°, 45°, 90°, 135°, 180°$;
- area (in km^2): 0.1, 1, 10, 100, 1000, 1e4, 1e5, 1e7, 1e8, 5.1e8;
- time (in days): 1, 7, 30, 90, 180, 365.25, 730.5, 1826.25, 3652.5.

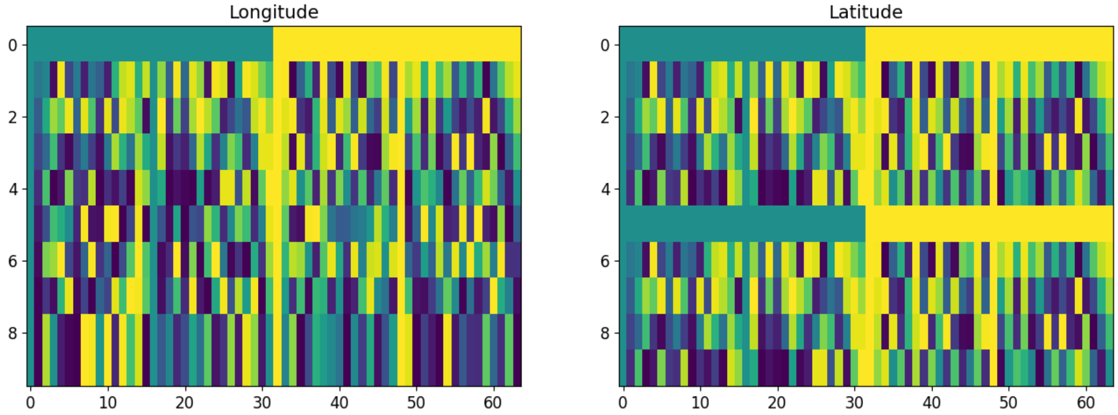


Figure 11. Fourier encoding visualization for geolocation (longitudes and latitudes).

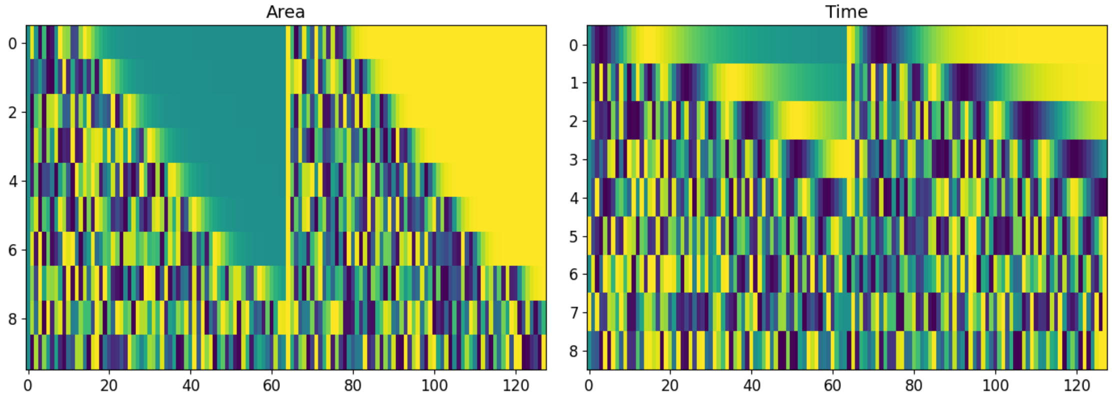


Figure 12. Fourier encoding visualization for area (left) and time (right).

Ablation on metadata dropping ratio In practice, metadata is not always available as input. We thus randomly drop part of the metadata during pretraining, and use learnable metadata tokens to fill missing metadata encodings. To choose the best metadata dropping probability, Tab. 6 conducts a corresponding ablation study, where we perform k-NN evaluation on three image classification tasks. The table suggests that a relatively high dropping ratio helps improve the model's performance.

Table 6. Ablation study on the dropping ratio of metadata. We report overall accuracy with k-NN evaluation.

	EuroSAT-S1	EuroSAT-S2	EuroSAT-RGB
metadata (drop 0.1)	77.8	88.7	79.9
metadata (drop 0.3)	73.7	86.3	77.5
metadata (drop 0.5)	77.8	**89.6**	78.7
metadata (drop 0.7)	**81.0**	89.5	**78.9**
metadata (drop 0.9)	78.5	88.2	74.8

Ablation on metadata details Moving further, we wonder how much benefit each metadata component brings to the model, as well as how the format of each metadata will affect the performance. To answer these questions, Tab. 7 conducts additional ablation on specific metadata components. Results show that geolocation gives the most significant improvement, followed by area and time. Interestingly, geographic coordinates perform better than Cartesian coordinates despite their distortion in high-latitude regions. Using the area corresponding to the true surface coverage (e.g., cropping and resizing make the true surface coverage smaller) is necessary, without which the performance begins to drop. Using the day of the year and absolute days above one-year-period perform similarly. We use the latter such that it's convenient to extend to long time series in the future.

Table 7. Ablation study on the benefits of each metadata type. We report overall accuracy with k-NN evaluation. Gray rows are alternative formatting options for the metadata. Performance increases/decreases are compared to the best formatting option of previous metadata.

	EuroSAT-S1	EuroSAT-S2	EuroSAT-RGB
no metadata	56.9	88.3	70.1
+ location (x,y,z)	75.8 ↑ 18.9	88.7 ↑ 0.5	73.3 ↑ 2.8
/+ location (lon,lat)	78.2 ↑ 21.3	88.7 ↑ 0.4	76.5 ↑ 6.5
+ area (raw)	77.8 ↓ 0.4	88.1 ↓ 0.6	73.7 ↓ 2.8
/+ area (aug)	80.3 ↑ 2.2	89.3 ↑ 0.6	77.4 ↑ 0.8
+ time (dayofyear)	80.0 ↓ 0.3	89.5 ↑ 0.2	78.9 ↑ 1.5
/+ time (absolute)	81.0 ↑ 0.7	89.5 ↑ 0.2	78.9 ↑ 1.5

B.5. Pretraining details

Data We pretrain Copernicus-FM on the joint 220K-grid subset of Copernicus-Pretrain, with each grid being one sample unit containing aligned images from all eight modalities. For fast data loading, we convert the raw dataset into webdataset[1] format, with one grid cell being one minimum sample in the shards. During training, one image from each modality is sampled from one grid cell to construct the input for each iteration. For S1/2, we normalize the image values with channel-wise mean and standard deviation. For S3, we multiply each channel with its corresponding scale factor[2]. For S5P, we use the raw values, and replace NaN pixels with zero. For DEM, we divide the pixel values by 10000. We apply simple data augmentations to each modality, including random resized cropping with scale $[0.2, 1.0]$ to its corresponding input size and random horizontal flipping. Each image comes with its metadata, including geolocation (central coordinates in lon/lat), patch area (calculated from GSD and patch size in km^2), and time (number of days since a reference date 1970-01-01). The geolocation and patch area are adapted dynamically based on the cropping parameters in data augmentation. Note that despite this adaptation, due to geographical projection the patch area doesn't strictly reflect the surface area, but is accurate enough for our pretraining purpose. While S1/2/3 images have exact acquisition dates, S5P images are monthly mean and DEM doesn't have a specific acquisition date. Therefore, we use the first day of the month for one S5P image, and the first day of the year 2015 for all DEM images.

Model We use a standard vision Transformer [10] for the core backbone—e.g., a ViT-Base has 768 hidden dimensions, 12 Transformer blocks, and 12 attention heads. The MLP and attention architectures for the spectral and variable hypernetworks are identical to Xiong et al. [30]. For the light decoder to conduct masked image modeling (MIM) pretraining, we also follow Xiong et al. [30] and He et al. [14] with 512 hidden dimensions, 8 Transformer blocks, and 16 attention heads. For continual distillation, a projector is used to project the output feature from the student to the frozen teacher model, both after global average pooling.

Loss We conduct MIM and continual distillation for pretraining. For MIM, we generally follow He et al. [14] to reconstruct masked-out patches for each input modality. The masking ratio is 70% for all modalities following previous performance studies of MIM in EO [26, 28]. For distillation, we distill RGB channels of S2 from frozen DINOv2 [21] (ViT-Base with patch size 14) with loss weight 0.1, and full channels of S1 and S2 from frozen SoftCon [27] (ViT-Base with patch size 14) with loss weight 0.2. The former serves as an anchor to control the latent space with general vision knowledge, such that

[1] https://github.com/webdataset/webdataset

[2] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S3_OLCI

the model can be used on high-resolution or RGB data despite only being pretrained on medium to low resolution Sentinel images. The latter serves as an accelerator to make training converge faster, as well as offering global representation guidance complementary to the main MIM objective. Our preliminary experiments suggest the benefits of the latter S1/2 distillation decrease with longer training times and larger models.

Training We pretrain Copernicus-FM on 220K Copernicus-Pretrain grids for 100 epochs. The effective batch size is 288. The basic learning rate is 1.5e-4 for batch size 256, and is linearly scaled for varied batch sizes. We warm up the learning rate for 10 epochs, and then apply a cosine decay schedule. We use the AdamW optimizer, with a weight decay of 0.05. One training run takes 512 GPU hours on NVIDIA A100 GPUs, or 128 node hours on one compute node with 4 A100 (40GB).

C. Copernicus-Bench

This section presents curation details, more characteristics, and additional visualizations for datasets within Copernicus-Bench, as well as implementation details for the benchmark.

C.1. Comparison to existing EO benchmarks

Tab. 8 shows a detailed comparison between Copernicus-Bench and several existing EO benchmarks.

Table 8. A comparison of existing EO benchmarks.

	# tasks	task types	modalities	resolution	task range
SustainBench [31]	15	cls, seg, reg	RGB, MS	0.6–30 m	surface
GEO-Bench [17]	12	cls, seg	RGB, MS, HS, SAR	0.1–15 m	surface
FoMo-Bench [5]	16	cls, seg, obj	RGB, MS, HS, SAR	0.01–60 m	surface
PhilEO Bench [11]	3	seg, reg	MS	10 m	surface
Copernicus-Bench (ours)	15	cls, seg, reg, cd	MS, SAR, atmos. var.	10–1000 m	surface, atmosphere

C.2. Benchmark curation

Copernicus-Bench consists of 15 datasets organized into 3 levels of tasks covering all primary Copernicus Sentinel missions. Among them, nine are derived from existing datasets with permissive licenses, and six are newly curated to fill in the gaps of ML-ready datasets for S3/5P sensors.

Sourced datasets Nine out of 15 datasets in Copernicus-Bench are extracted or adapted from existing datasets:

- **Cloud-S2**: This is a multi-class cloud segmentation dataset derived from CloudSEN12+ [1], one of the largest Sentinel-2 cloud and cloud shadow detection datasets with expert-labeled pixels. We take 25% samples with high-quality labels, and split them into 1699/567/551 train/val/test subsets.
- **EuroSAT-S1 and EuroSAT-S2**: These two are multi-class land use/land cover classification datasets taken from EuroSAT [15] and EuroSAT-SAR [28]. We follow the train/val/test splits defined in Neumann et al. [20] with 16200/5400/5400 train/val/test images. Images of the two datasets are one-to-one paired, thus they can also be combined to serve as a multimodal image classification dataset. These two datasets do not have time metadata.
- **BigEarthNet-S1 and BigEarthNet-S2**: These two datasets are sourced from BigEarthNet-v2 [9], a large-scale S1/2 dataset for multilabel land use/land cover classification. We sample a 5% subset (11894/6117/5991 images) from each of the official train/val/test splits, respectively. Images from the two datasets are again one-to-one paired, thus they can be combined to serve as a multimodal multilabel image classification dataset. In addition, each S1/2 image pair has a corresponding land cover map in 100 m resolution, thus they can also be used as pixel-level segmentation datasets.
- **DFC2020-S1 and DFC2020-S2**: These two are land use/land cover segmentation datasets derived from the IEEE GRSS Data Fusion Contest 2020 (DFC2020) [13]. We take S1/2 images and 10 m-resolution labels from the original test set, and further split them into 3156/986/986 train/val/test subsets. Again, images from S1 and S2 datasets are one-to-one paired, thus they can be combined to serve as a multimodal semantic segmentation dataset. These two datasets do not have geolocation and time metadata.
- **Flood-S1**: This is a flood segmentation dataset extracted from a large flood mapping dataset Kuro Siwo [4]. The original dataset is organized according to various flooding events around the globe. We take a random subset of samples that contain at least the water class to construct 3000/1000/1000 train/val/test subsets. Each sample contains two pre- and one post-event S1 SAR image, forming a time-series segmentation or a change detection dataset. By default, we use one pre-event and one post-event image in Copernicus-Bench.
- **LCZ-S2**: This is a multi-class scene classification dataset derived from So2Sat-LCZ42 [32], a large-scale local climate zone classification dataset. We randomly select 25K samples from the training set of the ”cultural-10” version to construct new 15000/5000/5000 train/val/test subsets. The original data contains also S1 data, thus this dataset can also be extended to an S1 task and a multimodal task. This dataset does not have geolocation and time metadata.

New datasets Six of 15 datasets in Copernicus-Bench are newly curated:

- **Cloud-S3**: This is a cloud segmentation dataset with raw images from Sentinel-3 OLCI and labels from the IdePix [29] classification algorithm. We first download a few large cloudy S3 tiles (about 4800 × 400 pixels) distributed across the globe, and then apply the IdePix algorithm using the ESA SNAP toolbox to get multi-class cloud masks. After that, we manually check the quality of the generated masks, filter out low-quality tiles, and get seven big tiles with high-quality cloud labels. Next, we remap the label IDs, georeference the tiles to GeoTIFFs, and use GDAL to crop the large tiles into small patches with size 256 × 256 pixels. We remove boundary patches filled with NaN pixels, and split all high-quality patches into 1197/399/399 train/val/test subsets. The class names for the cloud masks are: invalid, clear, cloud-sure, cloud-ambiguous, cloud-shadow, and snow-ice, of which "invalid" should be ignored during training. Apart from the multi-class labels, for each image we also have one binary cloud mask. Therefore, the Cloud-S3 dataset can serve as both a multi-class and a binary segmentation dataset.
- **LC100Cls-S3 and LC100Seg-S3**: These two datasets are based on Sentinel-3 OLCI images and CGLS-LC100 [7] land cover maps. CGLS-LC100 is a product in the Copernicus Global Land Service (CGLS) portfolio and delivers a global 23-class land cover map at 100 m spatial resolution. We pick the map product for 2019, and sample and download S3 images and LC100 labels for about 10K locations across the globe using GEE. For each location, we download a land cover map with about 288 × 288 pixels, and four seasonal S3 OLCI images each with about 96 × 96 pixels (300 m resolution). Despite using bright pixel percentage to simulate cloud filtering, the resulting images still contain a large volume of clouds. To tackle this issue, we train a cloud detection model based on the previously introduced Cloud-S3 dataset and filter out model-detected cloudy images. After a final quality check, we get about 8K samples each with a land cover map and a time series of S3 images. We divide them into 5181/1727/1727 train/val/test subsets to construct the LC100Seg-S3 dataset. For LC100Cls-S3, we integrate multi-label annotations from the land cover maps for each sample, constructing a multilabel classification dataset. Note that the number of S3 time stamps for different samples may differ because of the cloud filtering process. Apart from the time series, we also pre-select one image for each sample, constructing the "static" version of LC100Cls-S3 and LC100Seg-S3, which is the default mode in Copernicus-Bench.
- **Biomass-S3**: This regression dataset is based on Sentinel-3 OLCI images and CCI biomass [22]. The biomass product is part of the European Space Agency's Climate Change Initiative (CCI) program and delivers global forest above-ground biomass at 100 m spatial resolution. We pick the product for 2020, and the layer of above ground biomass (AGB, unit: tons/ha, i.e. Mg/ha) as regression ground truth, which is defined as the mass, expressed as oven-dry weight of the woody parts (stem, bark, branches and twigs) of all living trees excluding stump and roots. We sample representative regions across the globe, and download corresponding S3 images (one for each season) from GEE and biomass maps from the CCI open data portal. We crop the S3 images into patches with about 96 × 96 pixels (300 m resolution), and the corresponding biomass maps into patches with about 288 × 288 pixels. Similar to LC100Cls-S3 and LC100Seg-S3, the resulting S3 images contain a large volume of clouds, thus we use again the cloud detection model to filter out cloudy images. After a final quality check, we acquire 5K samples each with a biomass map and a time series of S3 images. We divide them into 3000/1000/1000 train/val/test subsets to construct the Biomass-S3 dataset. Note that the number of S3 time stamps for different samples may differ because of the cloud filtering process. Apart from the time series, we also pre-select one image for each sample, constructing the "static" version of Biomass-S3, which is also the default mode in Copernicus-Bench.
- **AQ-NO2-S5P and AQ-O3-S5P**: These two regression datasets are based on Sentinel-5P NO_2 and O_3 images and EEA air quality data products [16]. The European Environment Agency (EEA) air quality product provides values for the human health related indicators of air pollutants at 1 km^2 grid covering the whole Europe, combining monitoring air quality data in a "regression-interpolation-merging-mapping" methodology and the observational values of the air quality monitoring stations used in the interpolation. We pick the products in 2021 for NO_2 (annual average concentration) and Ozone (O_3, 93.2 percentile of maximum daily 8-hour means, SOMO35) as regression ground truth. We sample and download S5P NO_2 ("tropospheric_NO2_column_number_density") and O_3 ("O3_column_number_density") images from GEE, and EEA NO_2 and O_3 maps from EEA datahub[3]. We use a sample patch size of about 56 × 56 pixels for both S5P and EEA. For S5P, we download two versions: 1) annual mean, and 2) seasonal mean for each season. After filtering out NaN patches and a final quality check, we get 1480/493/494 train/val/test samples for both NO_2 and O_3, each with an "annual" mode of 1 S5P image and a "seasonal" mode of 4 S5P images. "Annual" is the default mode in Copernicus-Bench.

C.3. Benchmark characteristics

Example visualization Figs. 13 to 22 visualize some examples for each dataset in Copernicus-Bench.

[3] https://www.eea.europa.eu/en/datahub

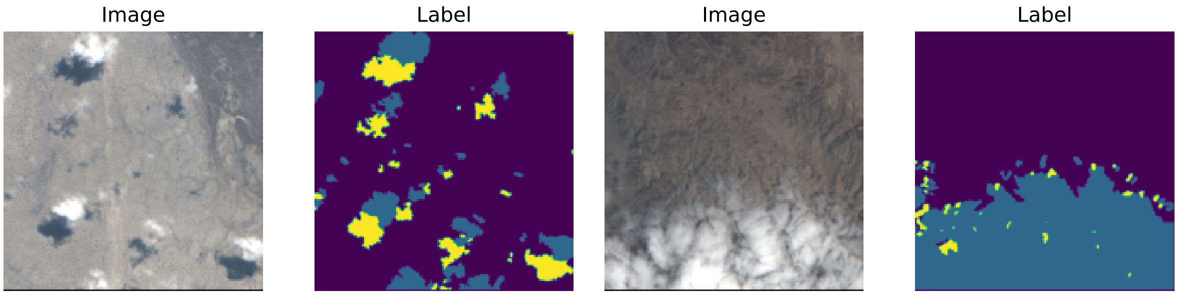


Figure 13. Copernicus-Bench-Cloud-S2.

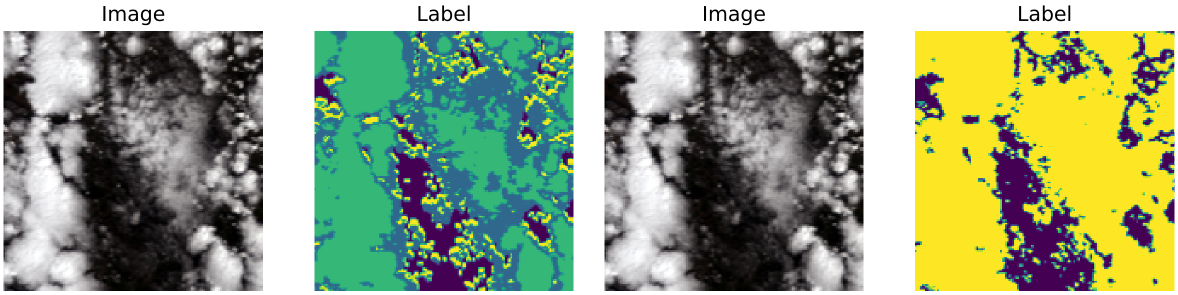
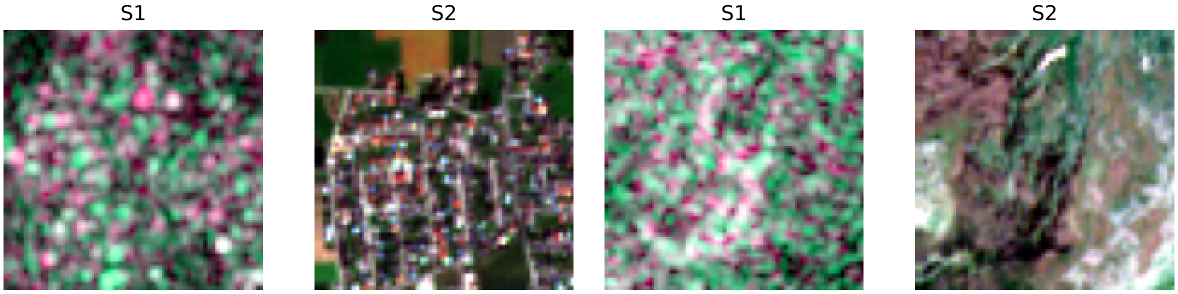


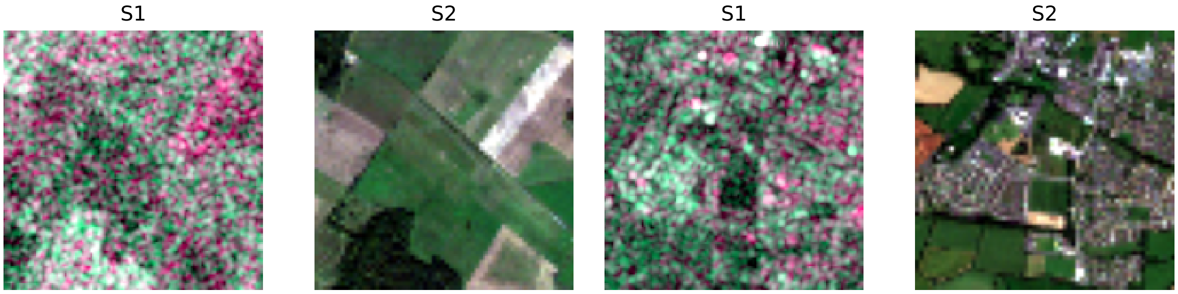
Figure 14. Copernicus-Bench-Cloud-S3. Left: “multi-class” mode. Right: “binary” mode.



Residential

HerbaceousVegetation

Figure 15. Copernicus-Bench-EuroSAT-S1 and Copernicus-Bench-EuroSAT-S2.



['Arable land', 'Pastures', 'Broad-leaved forest']

['Urban fabric', 'Arable land', 'Pastures']

Figure 16. Copernicus-Bench-BigEarth-S1 and Copernicus-Bench-BigEarth-S2.

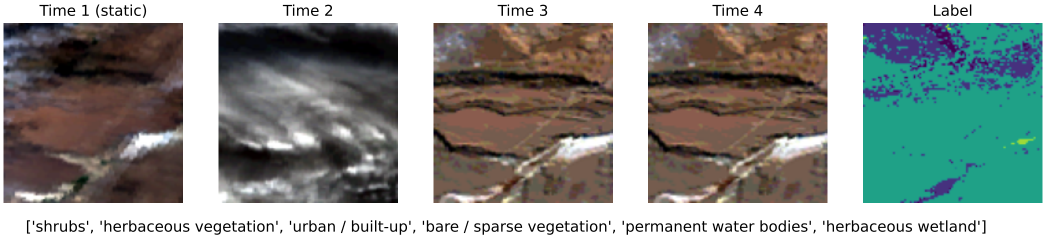

Time 1 (static)
Time 2
Time 3
Time 4
Label
['shrubs', 'herbaceous vegetation', 'urban / built-up', 'bare / sparse vegetation', 'permanent water bodies', 'herbaceous wetland']

Figure 17. Copernicus-Bench-LC100Cls-S3 and Copernicus-Bench-LC100Seg-S3. By default we pick one image per time series as "static" mode.

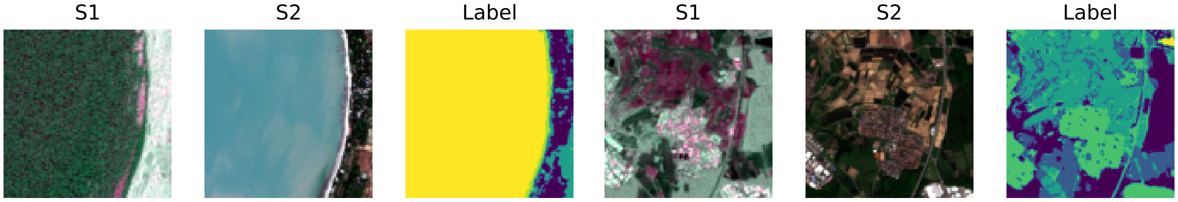


Figure 18. Copernicus-Bench-DFC2020-S1 and Copernicus-Bench-DFC2020-S2.

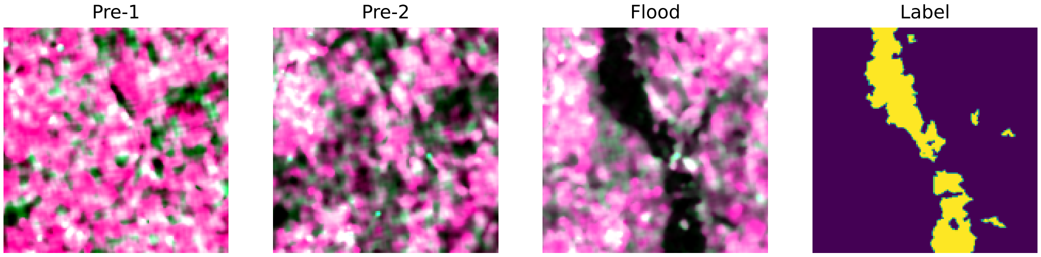


Figure 19. Copernicus-Bench-Flood-S1.

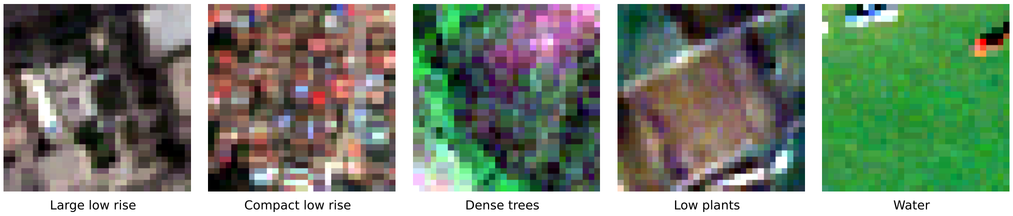


Figure 20. Copernicus-Bench-LCZ-S2.

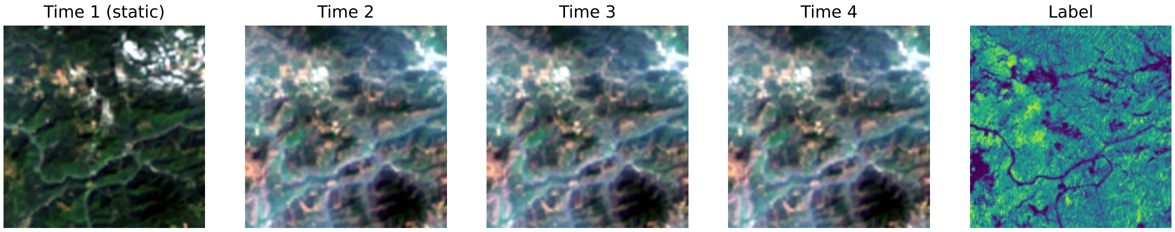


Figure 21. Copernicus-Bench-Biomass-S3. By default we pick one image per time series as "static" mode.

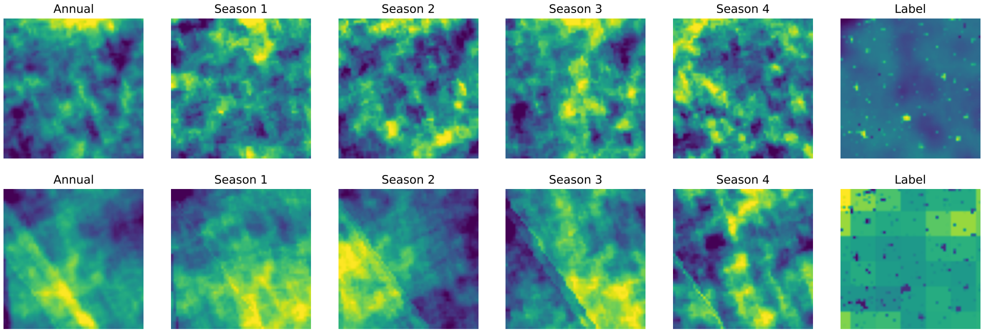


Figure 22. Copernicus-Bench-AQ-NO2-S5P and Copernicus-Bench-AQ-O3-S5P. By default we pick the “annual” mode.

Geographical distribution Fig. 23 illustrates the geographical distribution of datasets in Copernicus-Bench. Note that DFC2020-S1, DFC2020-S2, and LCZ-S2 do not have geolocation metadata.

Metadata information Tab. 9 lists the metadata information of the datasets in Copernicus-Bench.

Table 9. Copernicus-Bench metadata availability.

Level	Name	Task	Sensor	Bands	Location	Time	Area
L1	Cloud-S2	seg	S2 TOA	All 13 bands	✓	✓	✓
	Cloud-S3	seg	S3 OLCI	All 21 bands	✓	✓	✓
L2	EuroSAT-S1	cls	S1 GRD	VV, VH	✓	✗	✓
	EuroSAT-S2	cls	S2 TOA	All 13 bands	✓	✗	✓
	BigEarthNet-S1	cls	S1 GRD	VV, VH	✓	✓	✓
	BigEarthNet-S2	cls	S2 SR	12 bands (no B10)	✓	✓	✓
	LC100Cls-S3	cls	S3 OLCI	All 21 bands	✓	✓	✓
	DFC2020-S1	seg	S1 GRD	VV, VH	✗	✗	✓
	DFC2020-S2	seg	S2 TOA	All 13 bands	✗	✗	✓
	LC100Seg-S3	seg	S3 OLCI	All 21 bands	✓	✓	✓
L3	Flood-S1	cd	S1 GRD	VV, VH	✓	✓	✓
	LCZ-S2	cls	S2 TOA	10 bands (no B1, B9, B10)	✗	✗	✓
	Biomass-S3	reg	S3 OLCI	All 21 bands	✓	✓	✓
	AQ-NO2-S5P	reg	S5P NO2	tropospheric NO_2 column number density	✓	✓	✓
	AQ-O3-S5P	reg	S5P O3	O_3 column number density	✓	✓	✓

C.4. Benchmark implementation

We run all benchmark experiments on a single GPU, repeating three runs with different random seeds. We first benchmark two supervised baselines with ViT-S/16 and ViT-B/16, and then conduct frozen-encoder transfer learning for a set of pretrained models. For classification tasks, a linear layer is appended on top of the encoder; for segmentation and regression tasks, a UPerNet decoder with an auxiliary FCN decoder is appended on top of the encoder; for the flood segmentation task, we follow the segmentation design except that both pre- and post-event images are sent through the encoder and the difference features are sent to the decoder.

We use simplified data augmentations for training sets: horizontal and vertical flipping for classification tasks, and 90°-rotation, horizontal and vertical flipping for segmentation and regression tasks. No augmentation is used for validation and testing sets. Data normalization is performed on the input according to the pretrained model's preference. For most cases, normalization is performed by subtracting the channel-wise mean and dividing by standard deviation based on the pretrained-model-preferred statistics. If there is no preference, we recommend using the statistics calculated from the training set of

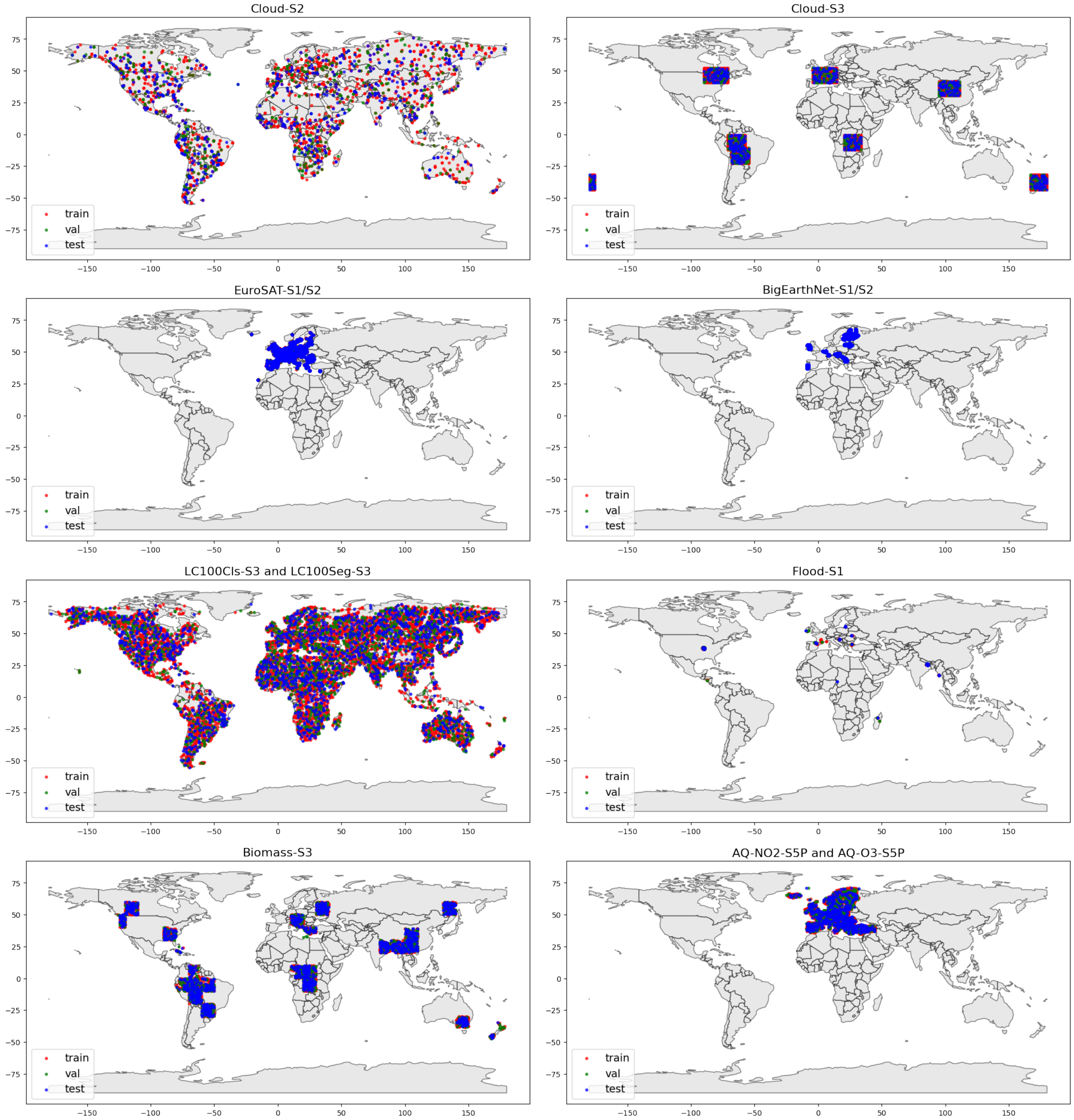


Figure 23. Geographical distribution of datasets in Copernicus-Bench.

each dataset as a standard. For regression tasks, we do mean/std (of the training set) normalization also on the targets to stabilize training. The predicted output is later converted back to the original scale to compute evaluation metrics.

We run each experiment for 50 epochs, and report the test set metrics based on the best validation scores. For classification tasks, we use a batch size of 64, the SGD optimizer, and cross entropy loss or multilabel soft margin loss for single-label or multi-label cases. For segmentation tasks, we use a batch size of 16, the AdamW optimizer, and cross entropy loss. For regression tasks, we use a batch size of 16, the AdamW optimizer, and L1 loss. Specially for air quality regression tasks (NO_2 and O_3), the targets may contain NaN pixels, thus we customize a masked L1 loss where the NaN pixels do not contribute to the loss calculation. For each model and dataset, we look for the best learning rate with a simple grid search from the pool [1e-4,1e-3,1e-2] (for AdamW) and [1e-2,1e-1, 1, 10] (for SGD). In most cases, the best learning rate is consistent across models but slightly varies across datasets.

D. Bridging EO and climate with grid embeddings

D.1. Climate prediction visualization

Figs. 24 and 26 visualize the prediction results on 10-year mean/std of the six climate parameters, comparing using the geocoordinates or a combination of coordinates and Copernicus-FM-generated grid embeddings as input data. Figs. 25 and 27 further plot the prediction error (Target-Prediction) of using only coordinates, coordinates and embeddings, and only embeddings. The figures show that using raw coordinates captures the general distribution of the climate parameters but tends to be over-smooth, while introducing EO-generated grid embeddings can capture finer details and extremes.

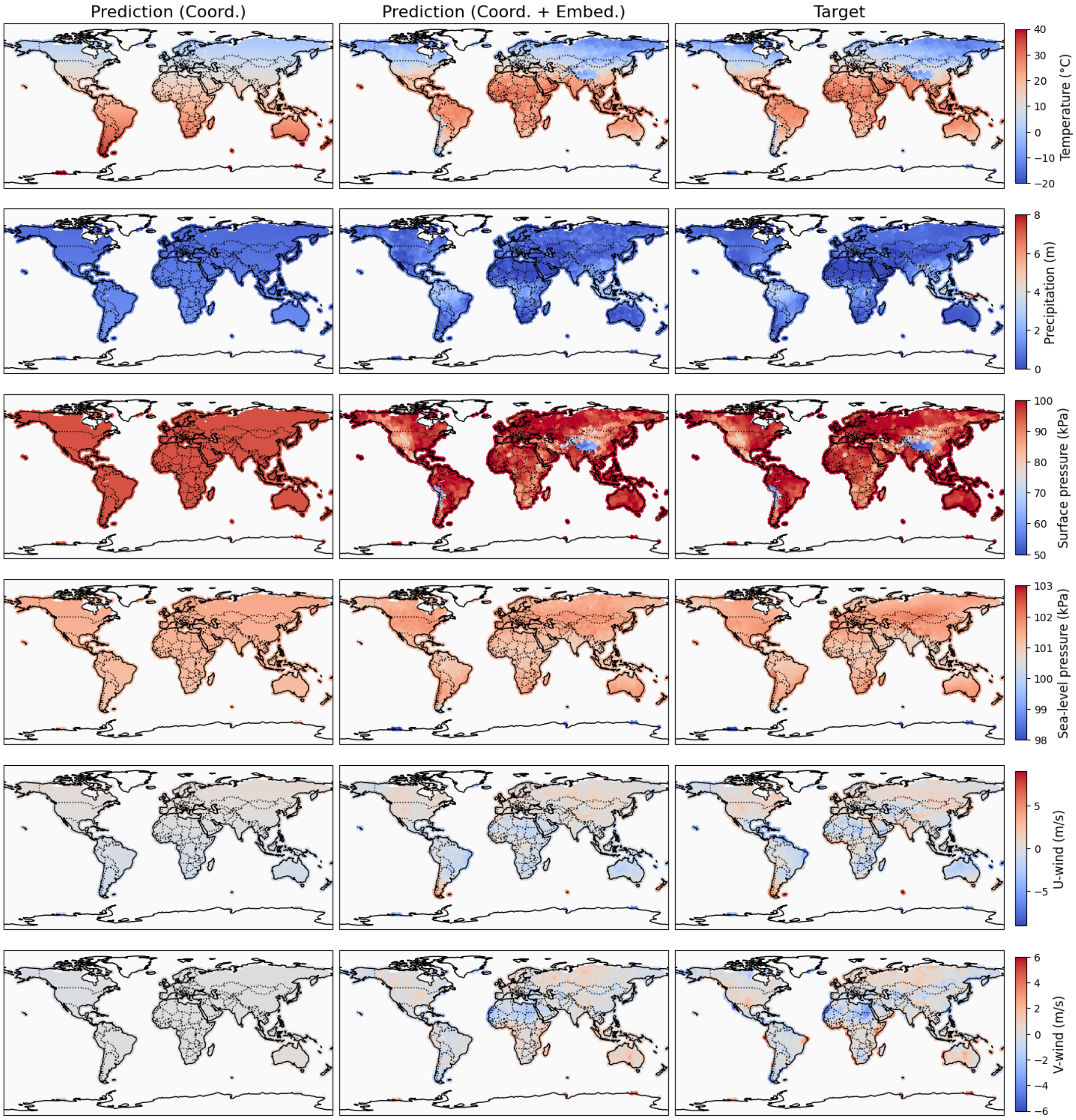


Figure 24. Visualization of climate prediction (10-year mean) results comparing different input sources.

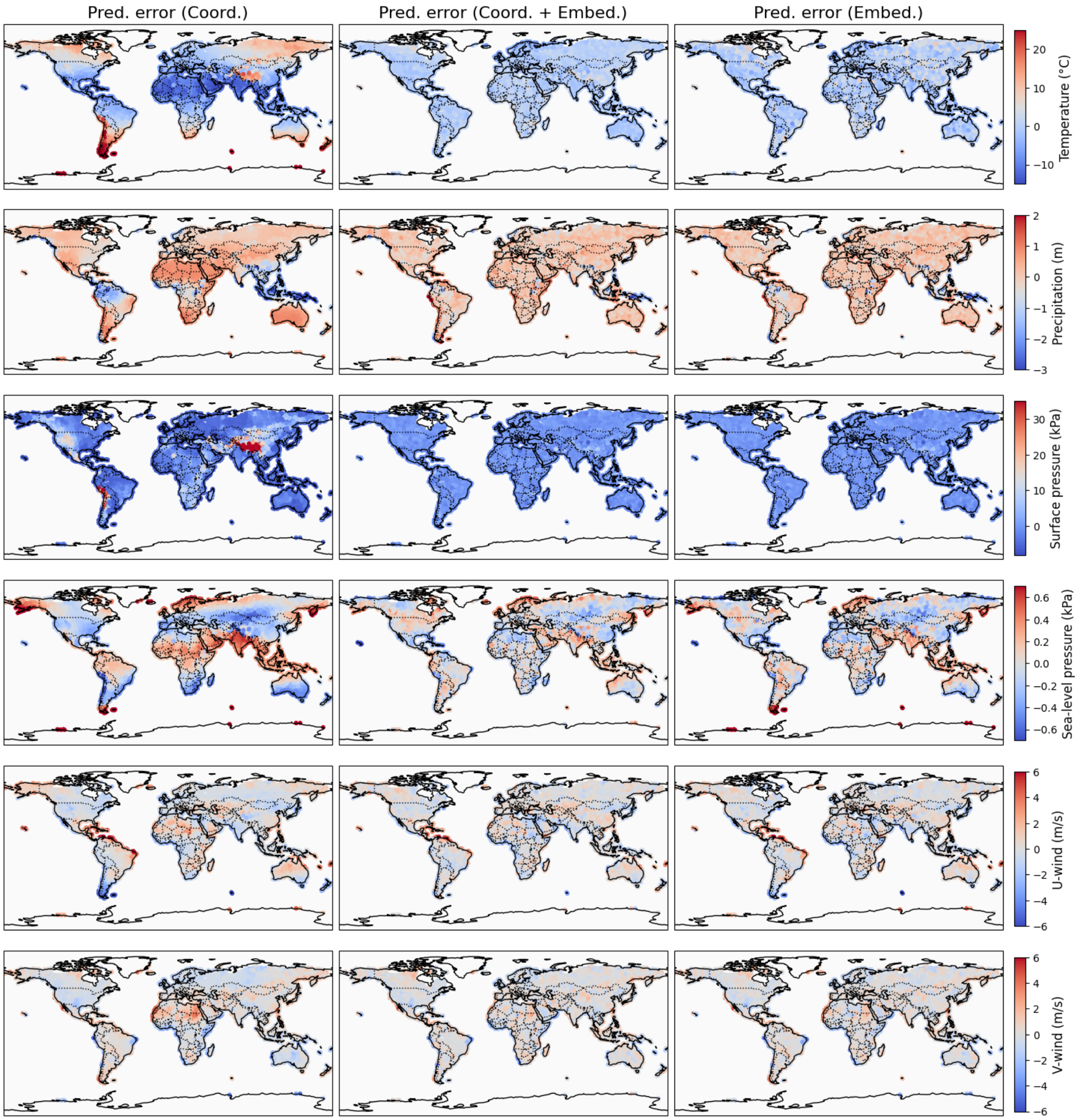


Figure 25. Visualization of climate prediction (10-year mean) errors comparing different input sources.

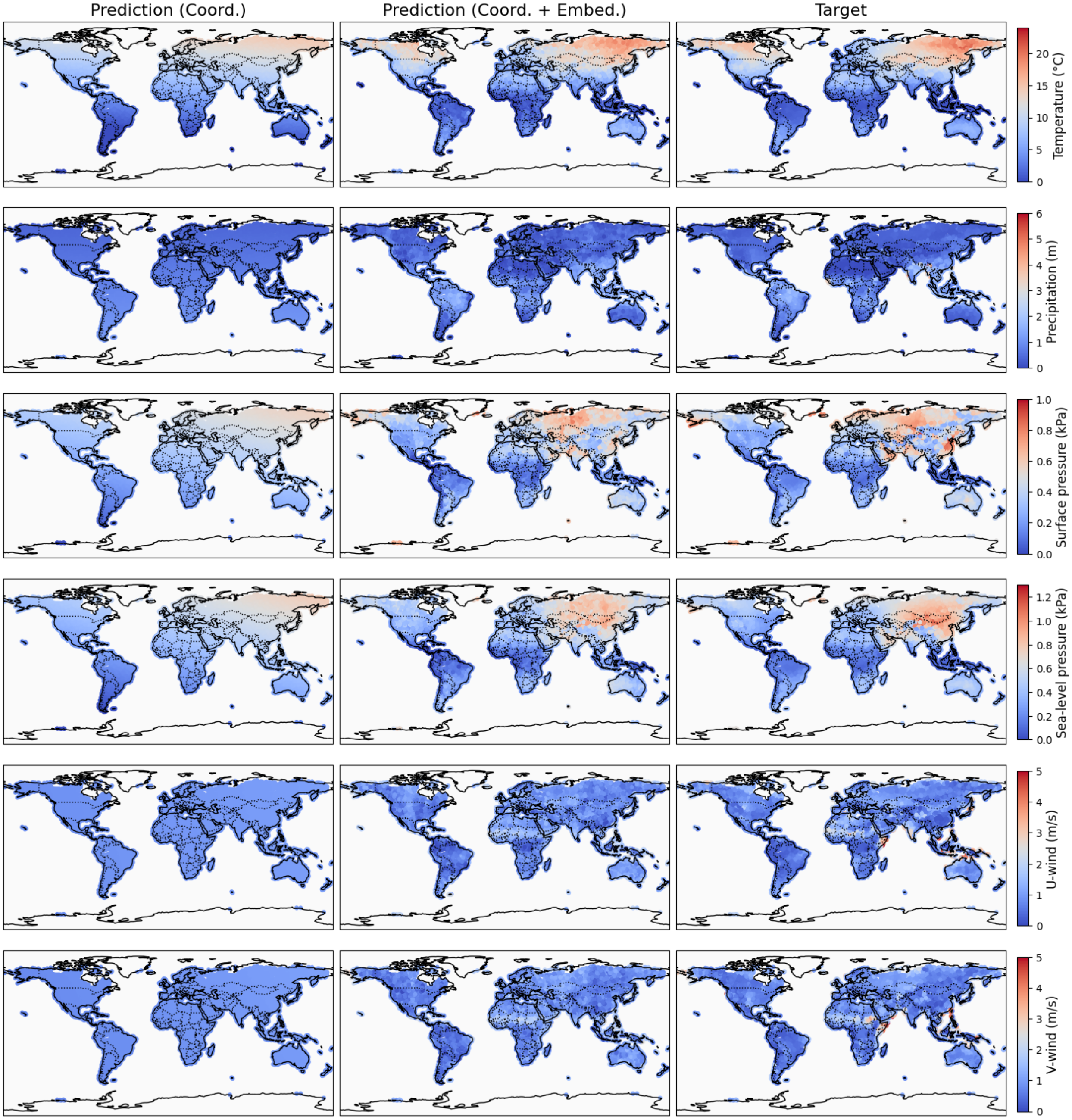


Figure 26. Visualization of climate prediction (10-year std) results comparing different input sources.

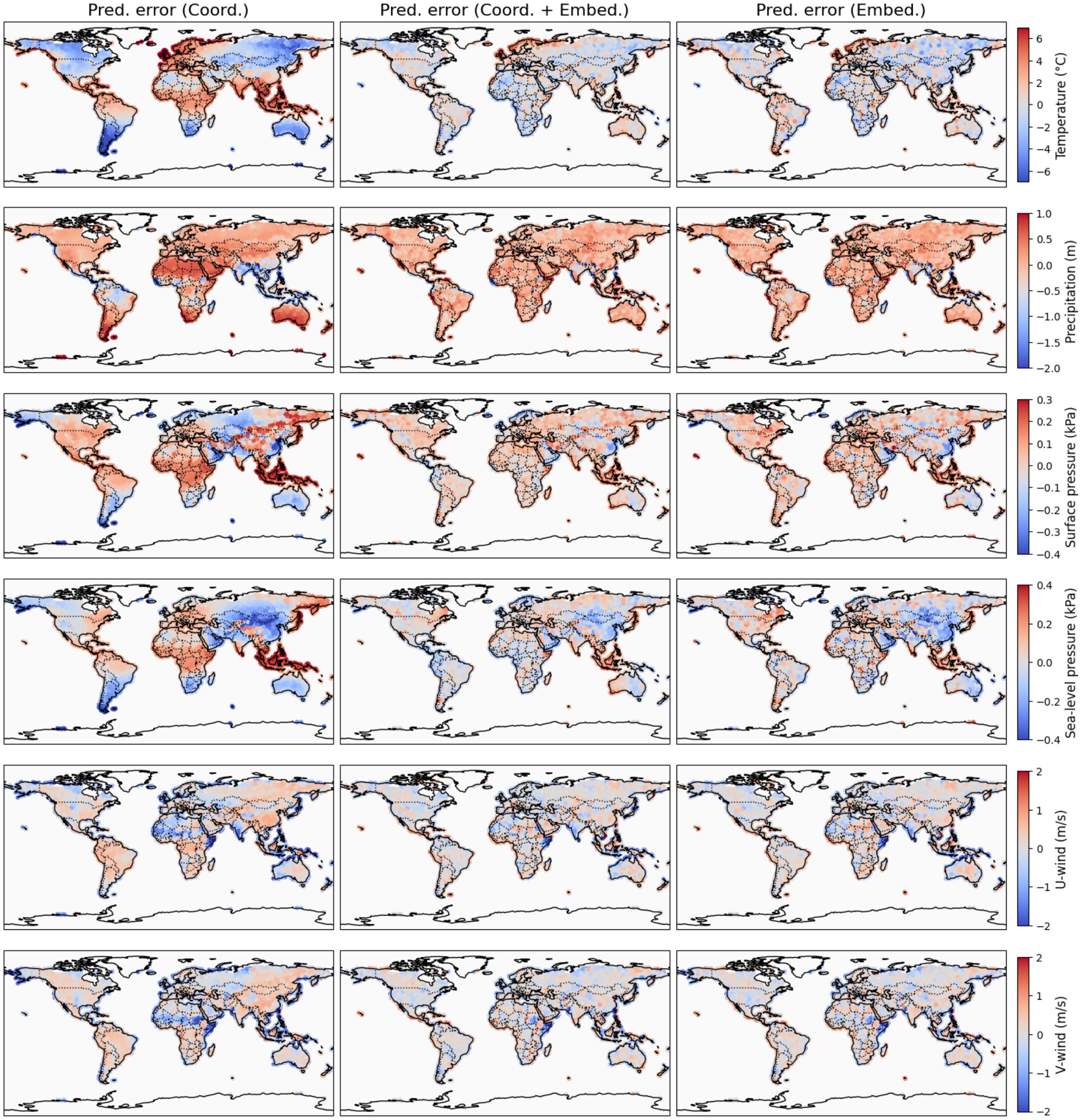


Figure 27. Visualization of climate prediction (10-year std) errors comparing different input sources.

D.2. Copernicus embedding dataset

Originally, 10K 0.25°×0.25° grids (with all 8 modalities) are sampled from the Copernicus-Pretrain dataset and encoded using the Copernicus-FM model to get image embeddings for each modality. The embeddings are averaged over different modalities to get one embedding vector for each grid, and later used for the climate prediction tasks to investigate the potential of bridging EO and climate. As a follow-up, we extend the embeddings to the whole globe using the full Copernicus-Pretrain dataset, constructing a ”global embedding map” at 0.25° with shape 721x1440x768 (filling ocean grids with 0). We term this embedding dataset **Copernicus-Embed-025deg**, which can be seen as a semantic map that integrates various sources of satellite observations at an extremely high compression ratio. This dataset makes it very convenient to link Earth's surface to the atmosphere (e.g., as improved static variables adding to ERA5), unlocking new possibilities in the development of weather/climate foundation models. Fig. 28 visualizes the Copernicus-Embed-025deg dataset with top-3 principal components as RGB channels.

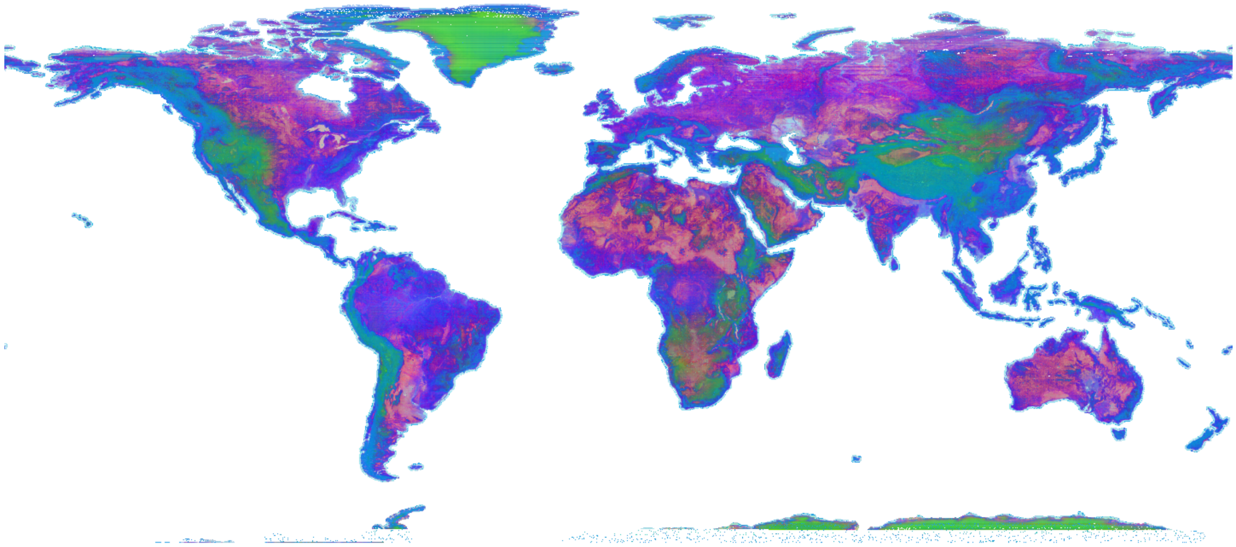


Figure 28. Visualization of the Copernicus-Embed-025deg dataset as a global embedding map (PCA to 3-dim).

E. License

All codes, datasets, and model weights will be publicly released under permissive licenses. All codes will be released on GitHub under the Apache-2.0 license, including the curation codes of Copernicus-Pretrain and Copernicus-Bench, the pretraining codes of Copernicus-FM, and the benchmarking codes for Copernicus-Bench. The Copernicus-Pretrain dataset, the newly-curated datasets in Copernicus-Bench, and the pretrained weights of Copernicus-FM will be released under the CC-BY-4.0 license, a copy of which will be hosted on public platforms like Hugging Face. We will also contribute our dataset, model, and benchmark to popular open-source libraries such as TorchGeo [25].

References

[1] Cesar Aybar, Lesly Bautista, David Montero, Julio Contreras, Daryl Ayala, Fernando Prudencio, Jhomira Loja, Luis Ysuhuaylas, Fernando Herrera, Karen Gonzales, et al. CloudSEN12+: The largest dataset of expert-labeled pixels for cloud and cloud shadow detection in Sentinel-2. *Data in Brief*, 56:110852, 2024. 10

[2] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 1

[3] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024. 5, 6

[4] Nikolaos Ioannis Bountos, Maria Sdraka, Angelos Zavras, Andreas Karavias, Ilektra Karasante, Themistocles Herekakis, Angeliki Thanasou, Dimitrios Michail, and Ioannis Papoutsis. Kuro Siwo: 33 billion m2 under the water. a global multi-temporal satellite dataset for rapid flood mapping. In *Advances in Neural Information Processing Systems*, pages 38105–38121, 2024. 10

[5] Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. FoMo-Bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *39th Annual AAAI Conference on Artificial Intelligence*, 2025. 10

[6] Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *arXiv preprint arXiv:2408.08447*, 2024. 1

[7] M. Buchhorn, B. Smets, L. Bertels, B. De Roo, M. Lesiv, N.-E. Tsendbazar, M. Herold, and S. Fritz. Copernicus global land service: Land cover 100m: collection 3: epoch 2019: Globe (version v3.0.1), 2020. 11

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 1

[9] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reBEN: Refined BigEarthNet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*, 2024. 10

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[11] Casper Fibaek, Luke Camilleri, Andreas Luyts, Nikolaos Dionelis, and Bertrand Le Saux. PhilEO Bench: Evaluating geo-spatial foundation models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2739–2744. IEEE, 2024. 10

[12] Alistair Francis and Mikolaj Czerkawski. Major TOM: Expandable datasets for Earth observation. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2935–2940. IEEE, 2024. 1

[13] Michael Schmitt; Lloyd Hughes; Pedram Ghamisi; Naoto Yokoya; Ronny Hänsch. 2020 IEEE GRSS data fusion contest, 2019. 10

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 8

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. 10

[16] Jan Horálek, Leona Vlasáková, Markéta Schreiberová, Nina Benešová, Philipp Schneider, Pavel Kurfürst, Frédéric Tognet, Jana Schovánková, Ondřej Vlček, Marta Garcia Vivanco, Mark Theobald, and Victoria Gil. ETC HE report 2023/3: Air quality maps of EEA member and cooperating countries for 2021. PM10, PM2.5, O3, NO2, NOx and BaP spatial estimates and their uncertainties. Technical report, European Environment Agency (EEA), 2024. Report provides air quality maps and exposure estimates for pollutants in EEA member and cooperating countries for 2021. 11

[17] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward foundation models for Earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023. 10

[18] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal Contrast: Unsupervised pretraining from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 1

[19] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multimodal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024. 1

[20] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 10

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 8

[22] M. Santoro and O. Cartus. ESA biomass climate change initiative (biomass_cci): Global datasets of forest above-ground biomass for the years 2010, 2015, 2016, 2017, 2018, 2019, 2020 and 2021, v5.01, 2024. 11

[23] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS–a curated dataset of georeferenced multispectral Sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 1

[24] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. SSL4EO-L: Datasets and foundation models for Landsat imagery. *Advances in Neural Information Processing Systems*, 36:59787–59807, 2023. 1

[25] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. TorchGeo: Deep learning with geospatial data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–12, Seattle, Washington, 2022. Association for Computing Machinery. 21

[26] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 1, 8

[27] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Multi-label guided soft contrastive learning for efficient Earth observation pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 8

[28] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 8, 10

[29] Jan Wevers, Dagmar Müller, Grit Kirches, Ralf Quast, and Carsten Brockmann. IdePix for Sentinel-3 OLCI algorithm theoretical basis document, 2022. 11

[30] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv e-prints*, pages arXiv–2403, 2024. 8

[31] Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. SustainBench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021. 10

[32] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, et al. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020. 10