THOR: A Versatile Foundation Model for Earth Observation Climate and Society Applications

Supplementary Material

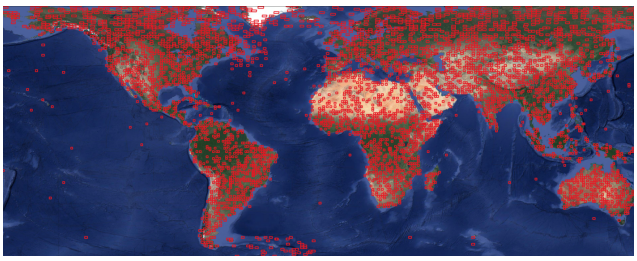


Figure S.1. Overview of THOR Pretrain sampled locations.

A. THOR Pretrain

The THOR FM is pre-trained on a new, diverse, and large-scale dataset named THOR Pretrain. This dataset is curated to learn representations that are robust to variations in global land cover, ocean phenomena, and cloud conditions.

THOR Pretrain unifies data from four major Copernicus Sentinel missions: Sentinel-1 SAR, Sentinel-2 MSI, Sentinel-3 OLCI, and Sentinel-3 SLSTR. These sensors provide diverse image modalities, including radar, multispectral and thermal sensors, with resolutions ranging from 10 m to 1000 m. In addition to the satellite data, the dataset includes a digital elevation model (DEM), diverse land cover maps, and ERA5-Land data. The dataset consists of 22TB of data from globally distributed locations (Fig. S.1).

A.1. Data, pre-processing and alignment

Rather than stacking millions of small image crops, we anchor our EO data sampling to the standard Sentinel-2 tile grid (110 $\times$ 110 km). For a given grid location and time, we sample the Sentinel-2 tile along with overlapping Sentinel-1 SAR, Sentinel-3 OLCI, and Sentinel-3 SLSTR data. Sentinel-3 data is selected from a 25 times larger area, centered at the Sentinel-2 tile, to account for its coarser resolution.

To ensure a diverse dataset of global land covers, ocean phenomena, and cloud conditions, we employ a stratified sampling strategy utilizing land cover and RGB maps of the world (see Sec. A.2 for details). This methodology is crucial to balance the dataset by actively prioritizing locations with high thematic and geographic diversity (e.g., [S1, S17]). A total of 6273 globally distributed locations were sampled (Fig. S.1).

A.1.1. Sensor data pre-processing

The Sentinel data are downloaded from Copernicus Data Space Ecosystem and preprocessed into netCDF files along with relevant metadata. The dataset consists of acquisitions spanning from January 1, 2016, to May 27, 2024.

Sentinel-1 SAR. Acquisitions include both ascending and descending passes, capturing available polarizations across modes (see Tab. S.5). The SAR data is processed to sigma-naught, corrected for thermal noise, geocoded using the Range Doppler algorithm. Two different resolutions of Sentinel-1 are constructed: 10 m and 60 m GSD. The 10 m GSD Sentinel-1 image is aligned with the corresponding Sentinel-2 data, whereas the 60 m GSD is processed to a larger area, bounded by the Sentinel-3 footprint.

Sentinel-2 MSI. Level 2A Sentinel-2 12 bands (see Tab. S.5) are collected into a single netCDF file along with metadata. The Scene Classification Map (SCL) product, which includes various land cover classes and a cloud mask, is also collected into the same netCDF file.

Sentinel-3 OLCI. Level 1 OLCI data are acquired. The top of atmosphere radiance (R_{TOA}) bands are converted to reflectance (L_{TOA}) using

$$R_{TOA}(\lambda) = \frac{\pi L_{TOA}(\lambda)}{E_0(\lambda) cos(\phi)}, \tag{1}$$

where E_0 is the solar spectral irradiance and ϕ is the sun zenith angle, both provided in the downloaded Sentinel-3 OLCI product file.

Further, the bands are resampled into the same UTM projection as the corresponding Sentinel-2 tile, but resampled to a GSD of 250 m and a geographic extent of 25 times larger area than the Sentinel-2 tile. This is done using the bilinear algorithm implemented in the *pyresample* Python library.

Sentinel-3 SLSTR. Level 1 SLSTR data are acquired. The Sentinel-3 SLSTR files are processed in the same manner as for OLCI: First, the top of atmosphere radiance bands are converted to reflectance using Eq. (1). Then the reflectance and brightness temperature bands are resampled to UTM projection and geographic extent similar to the OLCI product, except that the GSDs are 500 m and 1000 m for the reflectance and brightness temperature bands, respectively.

For SLSTR, cloud detection is performed using the SCDA version 2.0 algorithm [S13].

A.1.2. Auxiliary geospatial modalities (pretext targets)

The dataset also includes auxiliary geospatial modalities for reconstruction and prediction pretext tasks:

- Digital Elevation Model (pixel-level targets): DEM (extracted from Copernicus DEM GLO-30 and EU-DEM) are included, and the model reconstructs both slope and elevation at 10 m and 60 m GSD as part of the MAE reconstruction objective from Sentinel-1 and Sentinel 2 bands.
- Land cover maps (pixel-level targets): Several land cover products are incorporated to serve as pixel-level pretext tasks, accommodating the range of satellite sensors by varying in GSD from 10m to 500m. ESA WorldCover (10 m) [S23] and the Sentinel-2 SCL map is predicted from the Sentinel-1 and Sentinel-2 bands, the ESA GlobCover (300 m) [S2] is predicted from the Sentinel-3 OLCI bands, and MOD12Q1 map (500 m) [S6] is predicted from the Sentinel-3 SLSTR bands.
- ERA5-Land (image-level targets): The dataset includes ERA5-Land data based on daily statistics, derived from hourly land variables aggregated daily at 0.1 degrees resolution (approximate 9 km grid spacing). We select a diverse set of 17 variables covering temperature, hydrological cycles, snow cover, and vegetation indices (detailed in Table S.1). This data is used for image-level prediction pretext tasks.

To qualitatively validate our alignment pipeline, Fig. S.2 visualizes a complete sample tuple from the dataset. This visualization highlights the extreme heterogeneity THOR must resolve: the model must reconcile fine-grained textural details from the Sentinel-2 and Sentinel-1 (10 m) inputs with the broad-scale climatic context provided by the Sentinel-3 sensors.

As illustrated by the bounding boxes (Fig. S.2), the dataset preserves the spatial hierarchy of the sensors. The Sentinel-3 inputs cover a spatial footprint 25 times larger than the Sentinel-2 anchor tile, ensuring that the model captures large-scale atmospheric and thermal gradients that would be imperceptible in a narrow field-of-view crop. The inclusion of aligned DEM and Land Cover maps further confirms that the model receives dense topographic and semantic supervision alongside the raw radiometric data.

A.2. Stratified sampling strategy

The global land cover is not homogeneous, but highly imbalanced. Because oceans cover over 70% of the globe, uniformly sampling Sentinel-2 tiles would result in a dataset heavily skewed toward ocean environments. Even if we only sample tiles covering land, we will get a bias towards forest, desert and shrublands. Since increasing the pretraining data diversity enhances SSL performance [S1], we need to capture the variation of the land cover and sample the Sentinel-2 tiles in a stratified manner.

A.2.1. Land cover stratification

First, we perform an ocean/land split, selecting 80% of the Sentinel-2 tile from land areas.

To capture diversity of land areas, the strategy is based on k-means clustering of extracted features [S11, S17]. We use two data-sources to extract features from : ESA WorldCover maps and ESA Sentinel-2 RGB composite for 2022. Each of them are treated independently.

- Feature extraction: For each tile location, we divide the corresponding image data (WorldCover and RGB composite) into 224×224 crops. For ESA WorldCover maps we create a histogram of the 11 classes from each of crop, using bin counts as the feature vector. For the ESA Sentinel-2 2022 RGB composite, we use an ImageNet pre-trained ViT-MAE model to create a 768-dimensional embedding vector for each 224×224 crop.
- Clustering and probability: K-means clustering (with 1000 clusters) is applied independently to both the WorldCover feature vectors and the RGB embedding vectors to group similar crops. The sampling probability for each tile location is determined as the inverse of its cluster size, emphasizing rarity. By defining the per-tile sampling probability as proportional to $1/N$ (where N is the total number of tiles in that cluster), the expected number of samples drawn from each cluster becomes proportional to $N \times (1/N) = 1$. This ensures that we extract an equal expected number of samples per cluster, achieving a uniform distribution across classes.
- Tile selection: Tile sampling probability is the average of all crop probabilities within the tile, resulting in two probabilities: one from WorldCover and one from the Sentinel-2 RGB composite.

A.2.2. Ocean data sampling

To ensure comprehensive coverage of phenomena in the ocean, sampling probabilities utilize various maps:

- World Bank Global Shipping Traffic Density maps are used to calculate the normalized density of ship traffic and oil and gas installations per Sentinel-2 tile.
- Areas with a higher probability of containing icebergs and sea ice are defined based on existing maps and observations (e.g., specific longitudes for sea ice, and two large regions in the southern hemisphere for icebergs).

A.3. Final location sampling routine

The final per-tile sampling probabilities are a weighted combination of the land and ocean diversity scores, with an 80/20 split between land and ocean tiles. The detailed stratification for land and ocean samples is shown in Table S.2.

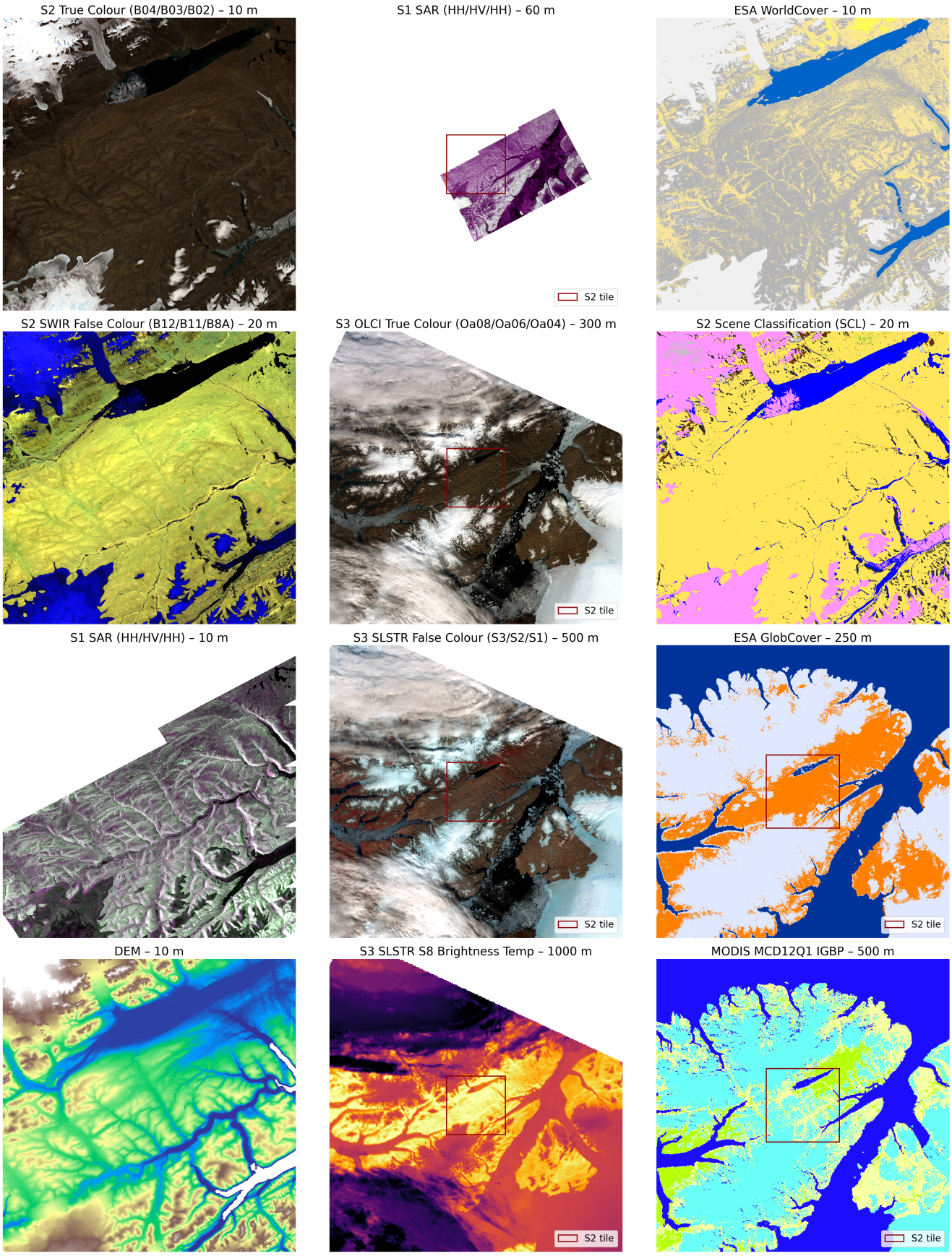


Figure S.2. Example images, from tile T19XDL on 2020-07-17. First column: 110x110km Sentinel-2 tile view of Sentinel-2 RGB bands and SWIR, Sentinel-1 (10m) and DEM, second column: Sentinel-1 (60m) and Sentinel-3 OLCI and SLSTR, third column: various land cover products. The footprint of the Sentinel-2 tile is shown as a red rectangle in larger views.

Table S.1. Description of ERA5-Land variables used to pre-train THOR, and included in THOR Pretrain.

Variable Name	Unit	Description
volumetric_soil_water_layer_1	None	Volumetric soil water fraction for layer 1 (0–7cm).
volumetric_soil_water_layer_4	None	Volumetric soil water fraction for layer 4 (100–289cm).
skin_temperature	K	Temperature of the surface of the Earth.
dewpoint_temperature_2m	K	Temperature at 2m to which air must be cooled for saturation.
temperature_2m	K	Air temperature at 2 meters above the surface.
soil_temperature_level_1	K	Soil temperature at layer 1 (0–7cm).
soil_temperature_level_4	K	Soil temperature at layer 4 (100–289cm).
snow_cover	None	Fraction of grid cell covered by snow.
snow_depth_water_equivalent	m	The depth of water that would result from melting the snow.
snowfall_sum	m	Accumulated snowfall (water equivalent).
snow_depth	m	Depth of the snowpack.
leaf_area_index_high_vegetation	None	Leaf area index fraction for high vegetation (e.g., trees).
leaf_area_index_low_vegetation	None	Leaf area index fraction for low vegetation (e.g., grass).
surface_pressure	Pa	Air pressure at the surface.
total_precipitation_sum	m	Accumulated total precipitation (rain and snow).
surface_runoff_sum	m	Accumulated water flowing over the land surface.
total_evaporation_sum	m	Accumulated evaporation from the surface.

Table S.2. Combined per-tile sampling probabilities

Category	Sub-category	Pct. of category
Land (80%)	Uniformly sampled from all land tiles	15%
	Sampled from ESA WorldCover diversity	75%
	Sampled from Sentinel-2 RGB composite	10%
Ocean (20%)	Uniformly sampled from all ocean tiles	5%
	Uniformly sampled from all coast tiles	40%
	Sampled from ship-density probabilities	10%
	Sampled from oil & gas installations-density	2%
	Uniformly sampled from sea-ice areas	30%
	Sampled from iceberg areas	13%

Table S.3. Sampling probabilities versus cloud coverage.

	Categories									
Cloud-cover interval [%]	<10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	>90
Sampling probability [%]	97.66	0.99	0.49	0.25	0.12	0.06	0.06	0.06	0.06	0.25

A.3.1. Temporal sampling

When a Sentinel-2 tile is sampled, the sampling routine selects an image from all available dates. While data constraints necessitate a balance between spatial and temporal coverage, the goal is to obtain an average of two images per tile. This is implemented by using a Poisson distribution with an expectation of one to determine the number of *additional* images to sample, ensuring at least one image per tile.

We generally aim to have as low cloud cover as possible, but since the model will encounter clouded images in inference, we want THOR Pretrain to contain clouded images as well. Hence, we assign sampling probabilities for each 10%-interval of cloud cover, and sample the image within (or as close as possible) to that interval (Tab. S.3).

A.4. Dataset summary

THOR Pretrain consists of a total of 18332 tile and date combinations, with 6273 unique Sentinel-2 tiles and 2926 unique dates, from 2016-01-01 to 2024-05-27.

Fig. S.3 illustrates the monthly distribution of the sampled observations, stratified by hemisphere. The distribution reveals two key characteristics of the dataset that align with the physical realities of optical remote sensing:

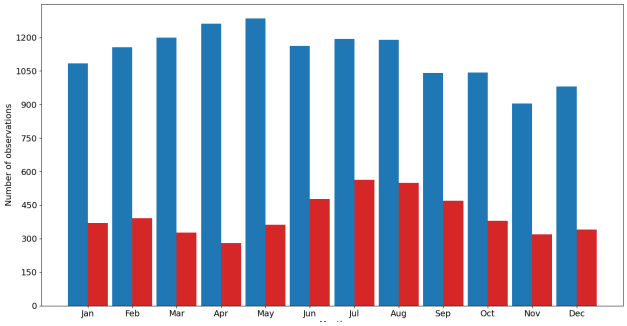


Figure S.3. Number of observations per month for northern (blue) and southern (red) hemisphere.

The total volume of samples from the Northern Hemisphere Fig. S.3 (blue bars) is consistently higher than that of the Southern Hemisphere (red bars). This reflects the Earth's geographical distribution, where approximately 68% of the global landmass resides in the Northern Hemisphere. Since our stratified sampling strategy prioritizes land tiles (80% land / 20% ocean split), the dataset naturally mirrors this global land distribution.

To validate the multi-modal density of THOR Pretrain, Tab. S.4 presents the co-occurrence matrix of all available sensor modalities. This distribution reveals three critical characteristics of the dataset that directly motivated our architectural choices:

- *High-volume multi-resolution alignment:* Approximately 10,000 overlapping samples between Sentinel-2 and Sentinel-3 (OLCI/SLSTR) bridge the 10 m – 1000 m resolution gap. This alignment enables the model to propagate fine-grained optical textures to coarse thermal and atmospheric readings.
- *Dense token supervision for radar-optical fusion:* Although 3,400 aligned Sentinel-1/Sentinel-2 pairs appear low in raw count, they represent full 110×110 km tiles rather than crops, yielding hundreds of millions of pixel-aligned tokens. Combined with stratified sampling for geodiversity, this provides a dense signal for learning radar-optical distributions without the redundancy of uncurated datasets.
- *Natural sparsity as a regularizer:* Variable sensor availability contrasts with the consistency of static auxiliary variables (land cover, DEM) across approximately 16,300 locations. This natural sparsity validates our independent per-band projection layers, acting as a regularizer that forces robustness to missing modalities and prevents over-reliance on single sensors.

It is important to note that we sample smaller crops from the full tiles during pre-training, i.e., Fig. S.2 is only an illustration of what modalities are available. During pre-training, random image locations in a given tile is sampled, and smaller crops of each available modality corresponding to the same footprint is extracted.

B. THOR foundation model implementation

B.1. Band groups

To handle the heterogeneous resolutions of the input sensors efficiently, we organize the input bands into 10 distinct groups as detailed in Tab. S.5. Grouping is primarily determined by the native GSD and the sensor source.

By grouping bands of identical resolution (e.g., Sentinel-2 10 m bands in Group 1, Sentinel-1 10 m bands in Groups 4 & 5), we allow the encoder to process each group with a patch number proportional to its information density. For instance, the thermal bands from Sentinel-3 (Group 10, 960 m GSD) require significantly fewer tokens than the optical bands from Sentinel-2 (Group 1, 10 m GSD) for the same input image footprint. This grouping strategy is fundamental to our token budget heuristic, ensuring that high-frequency spatial details are preserved where available, while minimizing computational waste on coarser modalities.

B.2. Multi-looking

Multi-looking is often applied in SAR applications to reduce speckle noise, a granular distortion inherent to coherent imaging systems like radar [S16]. By averaging independent "looks" (images) of the same scene, the random noise is smoothed out, which improves the image's radiometric quality at the expense of its spatial resolution. While aggregating the, say 10 m GRD pixels to 50 m, achieves a similar result in terms of reducing speckle and lowering resolution, it is technically not referred to as "multi-looking" in strict SAR processing terminology.

THOR has been pretrained using a random "multi-looking" by aggregating pixels to 10 m, 20 m, 30 m, 60 m, 120 m, 180 m or 240 m.

B.3. Model configurations

We train a family of THOR models ranging from Tiny to Large to evaluate scaling laws and deployment versatility. The specific architectural hyperparameters for each variant are provided in Tab. S.6. All models share the same unified encoder-decoder architecture but vary in embedding dimension, number of heads, and depth. Crucially, all variants support the dynamic input resolution (32^2 to 1024^2) and randomized patch sizes (4^2 to 32^2) described in the main text.

B.4. Token budget heuristic

Processing multi-modal data with randomized input sizes and patch sizes can lead to exploding sequence lengths if left unchecked. To address this, we implement a dynamic token budget heuristic, formally described in Algorithm 1.

Table S.4. Modality co-occurrence matrix (raw counts)

Modality	S2	S1:10m	S1:60m	S3:OLCI	S3:SLSTR	LC	DEM:10m	DEM:60m
S2	15310	3393	3528	9860	10952	14776	14556	14722
S1 10m	3393	4929	4896	2554	2919	4005	3966	3999
S1 60m	3528	4896	5121	2652	3032	4158	4099	4150
S3 OLCI	9860	2554	2652	11023	10574	10085	9927	10046
S3 SLSTR	10952	2919	3032	10574	12605	11366	11186	11321
LC	14776	4005	4158	10085	11366	16318	16095	16261
DEM 10m	14556	3966	4099	9927	11186	16095	16095	16094
DEM 60m	14722	3999	4150	10046	11321	16261	16094	16261

Table S.5. THOR input band grouping. Input bands are organized into 10 groups based on sensor source and spatial resolution. Note that Sentinel-1 data is split into coarse (60 m) and high-resolution (10 m) streams based on polarization/mode availability in the dataset. The Sentinel-1 IW and EW modes are mutually exclusive. † During training the GSD of the SAR is aggregated ("multi-looked") to 10, 20, 30, 60, 120, 180 or 240 m.

Group	Sensor	Bands	Default GSD (m)
1	Sentinel-2	Red, Green, Blue, NIR	10
2	Sentinel-2	RE1, RE2, RE3, RE4, SWIR1, SWIR2	20
3	Sentinel-2	CoastAerosol, WaterVapor	60
4	Sentinel-1	IW-VH, IW-VV, EW-VH, EW-VV	10/60†
5	Sentinel-1	IW-HV, IW-HH, EW-HV, EW-HH	10/60†
6	Sentinel-3 OLCI	Oa01, Oa02, Oa03, Oa04, Oa05, Oa06, Oa07	240
7	Sentinel-3 OLCI	Oa08, Oa09, Oa10, Oa11, Oa12, Oa13, Oa14	240
8	Sentinel-3 OLCI	Oa15, Oa16, Oa17, Oa18, Oa19, Oa20, Oa21	240
9	Sentinel-3 SLSTR	S1, S2, S3, S4, S5, S6 (reflectance)	480
10	Sentinel-3 SLSTR	S7, S8, S9 (thermal BT)	960

The algorithm operates by first sampling a global spatial footprint (C, C) in meters. For each band group g, we calculate number of tokens required based on a sampled patch size P_g and the GSD of the band group. If the cumulative number of tokens approaches the pre-defined maximum token budget, the algorithm dynamically adjusts the minimum allowable patch size for subsequent groups or caps the resolution. The ordering of the groups is randomly permuted to ensure the algorithm remains unbiased. This ensures that every training batch maximizes GPU utilization without causing Out-Of-Memory errors, regardless of the random footprint sampled.

B.5. Loss Details

The total loss $\mathcal{L}_{total}$ is a weighted sum of reconstruction, contrastive, and task-specific prediction losses. The specific weights (λ) assigned to each component are listed in Tab. S.7. We prioritize the reconstruction objective ($\lambda_1 = 1.5$) as it is the primary driver of feature learning in the MAE framework. The auxiliary tasks (ERA5, map prediction, orbital regression) are weighted lower (0.05 - 0.1) to act as regularizers and semantic guides without overwhelming the pixel-level reconstruction signal. The FFT loss [S9] is included with a small weight to stabilize high-frequency feature reconstruction.

C. Experiments

C.1. Extensive Pangaea results

We provide the complete tabulation of results for the Pangaea benchmark suite [S12] across three data availability regimes: 10% (Tab. S.8), 50% (Tab. S.9), and 100% (Tab. S.10). All THOR family model experiments are with patch size 6, input size of 108 and concatenation of the output features. These experiments validate that THOR provides good performance in low training data regimes.

In the 10% regime (Tab. S.8), THOR-B performs on par with the current state-of-the-art, TerraMind-B [S8]. This confirms that our flexible patching strategy, which allows for dense token representations at inference time, compensates for the lack of training labels by providing a richer signal to the decoder. For the full training dataset, Terra-

Table S.6. **THOR model family configurations.** Hyperparameters for the Tiny, Small, Base, and Large variants. All models support dynamic input resolutions and patch sizes (4^2 to 32^2) during pre-training. The Token budget is an approximate cap enforced during training to manage memory usage across heterogeneous inputs and is set to 1296 for all variants. The learning rate is set as base_lr $*$ (batch_size $*$ num_gpu) / 256.

Model	Layers	Embed dim	Heads	MLP ratio	Params	Base Training LR	Warmup epochs
THOR-T	12	192	3	4	∼7.6M	4e-4	10
THOR-S	12	384	6	4	∼25.8M	4e-4	10
THOR-B	12	768	12	4	∼94.1M	3e-4	20
THOR-L	24	1024	16	4	∼314.4M	3e-4	40

Algorithm 1 THOR dynamic token budget heuristic (ground-cover based)

Hyperparameters:
$B_{max} \leftarrow$ Maximum token budget (e.g., 1296)
$C_{range} \leftarrow [960, 46080]$ ▷ Ground-cover range (meters)
$P_{range} \leftarrow [P_{\min}, P_{\max}] = [4, 32]$ ▷ Patch size range (pixels)
$Groups \leftarrow$ List of sensor groups (e.g., [S1, S2, S3-OLCI, ...])

function SAMPLEPATCHPARAMETERS($Groups, C, B_{max}$)
▷ C is sampled: $C \sim \mathcal{U}(C_{range})$
    Randomly permute $Groups$ to get $(g_1, \ldots, g_G)$
    $T_{used} \leftarrow 0$
    for each group g in $(g_1, \ldots, g_G)$ **do**
        $H_g \leftarrow \left\lfloor \frac{C}{g.\text{GSD}} \right\rfloor; \quad W_g \leftarrow H_g$
        $B_{remain} \leftarrow B_{max} - T_{used}$
        if $B_{remain} \leq 0$ **then**
            break
        end if
        ▷ Token range as in implementation (2–32 grid limit)
        $T_{min} \leftarrow \left(\max\left(2,\ \lfloor H_g / P_{\max} \rfloor\right)\right)^2$
        $T_{max} \leftarrow \left(\min\left(32,\ \lceil H_g / P_{\min} \rceil\right)\right)^2$
        if $T_{min} > B_{remain}$ **then**
            continue ▷ Not enough budget for this group
        end if
        $T_{target} \leftarrow \min(T_{max}, B_{remain})$
        $G_{target} \leftarrow \sqrt{T_{target}}$ ▷ Target grid size per side
        $P_g \leftarrow \text{clip}\left(\left\lfloor \frac{H_g}{G_{target}} \right\rfloor,\ P_{\min},\ P_{\max}\right)$
        $T_{group} \leftarrow \left\lceil \frac{H_g}{P_g} \right\rceil \times \left\lceil \frac{W_g}{P_g} \right\rceil$
        $T_{used} \leftarrow T_{used} + T_{group}$
    end for
    return $\{(H_g, W_g, P_g) \mid g \in Groups$ with allocated budget$\}$
end function

Mind shows strong performance (achieving the top rank on average), but THOR-B remains highly competitive, outperforming the other models on PASTIS and CropMap tasks (Tab. S.10).

Feature aggregation strategy THOR processes input groups independently, requiring a fusion strategy to combine features before the decoder. We compare mean aggregation (averaging token embeddings across groups) against concatenation (stacking tokens along the channel dimen-

Table S.7. Loss function weights used in pre-training

Loss Component	Lambda	Weight Value
Reconstruction	λ_1	1.5
Contrastive	λ_2	0.1
ERA5	λ_4	0.1
Month	λ_5	0.1
Coordinates	λ_6	0.1
Orbit direction	λ_7	0.1
Incidence angle	λ_8	0.1
FFT	λ_9	0.01
Map prediction losses:		
SCL	$\lambda_{3,\text{SCL}}$	0.05
World Cover	$\lambda_{3,\text{WC}}$	0.1
Global Canopy	$\lambda_{3,\text{GC}}$	0.1
MCD12Q1	$\lambda_{3,\text{MCD}}$	0.1
DEM	$\lambda_{3,\text{DEM}}$	0.1

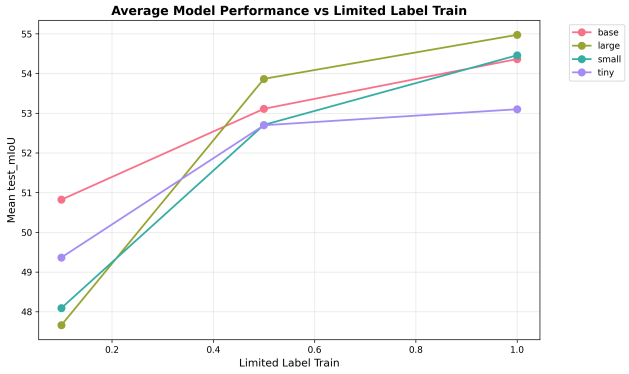


Figure S.4. Aggregated mIoU over all Pangaea benchmarks using 10, 50 and 100% training data for Tiny, Small, Base and Large model. Patch size 6, concat feature aggregation and input image size of 108 pixels.

sion). As shown in Tab. S.11, concatenation consistently outperforms mean aggregation, achieving 52.94% mIoU (vs. 49.90%) in the 10% data regime. This may suggest that distinct sensor modalities contain complementary, non-redundant information, and by averaging these features high-frequency modality-specific signals may be "washed out", whereas concatenation preserves the full feature variance often needed for fine-grained segmentation.

Validation of compute-adaptive patching A core premise of THOR is that smaller patch sizes yield denser feature maps, improving performance on pixel-level tasks. Tab. S.12 validates this hypothesis: reducing the patch size from 8 to 4 results in a significant performance boost, rising from 54.69% to 58.63% mIoU on the full dataset.

While smaller patches increase the sequence length (quadratic computational cost), they provide the necessary spatial granularity for segmentation tasks that coarse patches (e.g., 16×16) fail to resolve. This confirms that THOR's randomized patch pre-training successfully enables test-time adaptation to higher resolutions.

Scaling and data efficiency Fig. S.4 illustrates the scaling behavior of the THOR family (Tiny, Small, Base, Large) across data regimes. We observe a clear "crossover" effect: In data-scarce regimes (10%), the THOR-B model is the most robust performer. Notably, THOR-L underperforms on the 10% data (lowest starting point in Fig. S.4), indicating that massive models may be prone to overfitting when fine-tuning data is insufficient. In data-rich regimes (50-100%), THOR-L recovers and surpasses all other variants, validating standard scaling laws where capacity correlates with performance given sufficient supervision. However, the performance depends strongly on the dataset, as observed in Fig. S.5, where we show per-dataset performance of each model in the THOR family.

To further investigate the trade-off between computational cost and downstream performance, we conducted a series of experiments on four single-date Pangaea benchmarks using 10% of the training data. We compared the performance of a standard UPerNet decoder against a lightweight linear probe decoder across varying patch sizes.

As illustrated in Figs. S.6a, S.6c, S.6e, and S.6g, while the UPerNet decoder yields a performance boost in certain configurations, the linear decoder achieves competitive accuracy levels that are frequently on par with the much larger architecture. Critically, when analyzing the computational burden (Figs. S.6b, S.6d, and S.6f), the advantage of the linear approach becomes clear. As detailed in Tab. S.14, the UPerNet architecture requires approximately $1000\times$ more parameters than the linear decoder. This drastic reduction in decoder parameter count validates THOR as a true foundation model. The ability of a simple linear probe to match a complex non-linear decoder indicates that the pre-trained encoder produces highly semantic, linearly separable features. It suggests that in data-limited regimes, the heavy UPerNet decoder is largely redundant and potentially prone to overfitting, whereas THOR's dense representations may be deployed with minimal adaptation.

C.2. GEO-Bench image classification

We provide full results for the GEO-Bench [S10] image classification tasks (Tab. S.13). In the experiments, we have used a patch size of four and image size selected according to the minimum viable images [S19].

Table S.8. Extended Pangaea results with 10% training data in mIoU. Bold/underline mark best/second-best per column. Note: The shift in average rank values compared to Tab. 1 in the main paper is a mathematical artifact of the PANGAEA evaluation protocol; introducing additional models (THOR-S and THOR-L) to the pool re-calibrates the relative rankings for all methods.

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	FBP	DynEarthNet	CropMap	SN7	AI4Farms	Avg. Rank
CROMA	76.44	32.44	32.80	87.22	37.39	36.08	36.77	42.15	38.48	7.44
DOFA	71.98	23.77	27.68	82.84	27.82	**39.15**	29.91	46.10	27.74	11.78
GFM-Swin	67.23	28.19	21.47	62.57	55.58	28.16	27.21	39.48	32.88	14.11
Prithvi	77.73	21.24	33.56	86.28	29.98	32.28	27.71	36.78	35.04	11.67
RemoteCLIP	69.40	20.57	17.19	62.22	56.23	34.43	19.86	43.11	23.85	13.89
SatlasNet	74.79	29.87	16.76	83.92	37.86	34.64	29.08	49.78	13.91	12.11
Scale-MAE	75.47	21.47	22.86	64.74	48.75	35.27	13.44	49.68	26.66	12.33
SpectralGPT	**83.35**	20.29	34.53	83.12	39.51	35.33	31.06	36.31	37.35	10.00
S12-MoCo	73.11	19.47	32.51	79.58	35.57	32.24	36.54	49.46	37.97	12.44
S12-DINO	75.93	23.47	36.62	84.95	34.63	32.78	38.44	41.15	37.91	9.78
S12-MAE	76.60	18.44	31.06	84.81	35.56	30.59	35.29	40.51	23.60	13.00
S12-Data2Vec	74.38	17.86	33.09	81.91	37.27	33.63	34.11	40.66	22.85	13.89
TerraMind-B	77.39	**44.06**	**39.96**	84.43	54.00	37.35	35.65	43.21	38.59	**4.56**
UNet Baseline	79.46	24.30	29.53	**88.55**	52.58	35.59	13.88	46.08	34.84	8.11
ViT Baseline	75.92	10.18	38.44	81.85	**56.53**	35.39	27.76	36.01	**39.20**	10.11
THOR-T	75.98	41.65	36.26	82.70	42.81	34.03	37.82	58.52	38.56	7.11
THOR-S	77.29	42.64	38.48	84.21	42.81	35.31	40.39	**59.41**	12.31	6.56
THOR-B	76.90	40.67	38.93	86.29	42.80	35.21	**42.23**	55.93	38.90	4.67
THOR-L	75.57	36.43	39.21	87.34	43.51	36.10	36.77	55.79	18.26	6.44

Table S.9. Extended Pangaea results with 50% training data in mIoU. Bold/underline mark best/second-best per column.

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	FBP	DynEarthNet	CropMap	SN7	AI4Farms	Avg. Rank
CROMA	81.52	57.68	32.33	90.57	48.01	38.30	42.20	59.31	28.19	**5.11**
DOFA	78.02	55.21	28.60	88.39	36.90	**39.20**	30.93	47.06	26.69	11.78
GFM-Swin	74.36	**63.37**	20.41	71.61	63.14	31.25	31.42	59.83	28.43	10.33
Prithvi	80.89	40.79	33.13	89.69	40.27	33.43	42.51	49.45	29.27	9.33
RemoteCLIP	74.28	53.26	17.46	71.67	**65.92**	30.91	36.3	50.83	25.11	13.11
SatlasNet	75.97	52.24	16.78	89.45	46.04	36.34	35.29	60.74	27.08	10.00
Scale-MAE	75.47	46.87	23.26	72.54	62.11	32.60	20.32	**61.24**	26.40	12.33
SpectralGPT	76.40	58.00	34.61	87.52	21.71	36.52	32.09	56.28	27.46	10.56
S12-MoCo	79.79	42.90	32.59	89.22	46.92	34.45	41.32	56.21	28.38	8.89
S12-DINO	80.12	40.42	35.71	88.93	44.85	32.76	31.13	55.14	25.68	12.33
S12-MAE	80.13	44.29	31.15	88.43	45.63	33.29	28.07	55.55	27.50	11.44
S12-Data2Vec	79.82	41.22	33.42	86.58	46.73	32.61	28.53	56.94	25.84	11.89
UNet Baseline	**82.39**	43.87	30.25	**90.91**	55.42	35.14	36.30	46.82	**45.02**	7.89
ViT Baseline	78.17	28.77	38.71	86.08	57.32	37.33	39.53	49.21	38.37	9.00
THOR-T	78.22	53.87	36.40	89.29	45.20	35.00	48.58	60.03	27.67	7.44
THOR-S	79.14	52.14	38.20	90.42	44.99	36.64	45.41	59.46	27.95	7.11
THOR-B	79.15	49.43	39.50	89.05	45.39	36.66	50.81	60.24	27.76	6.67
THOR-L	76.69	52.30	**39.96**	89.92	46.03	37.36	**54.88**	59.67	27.96	5.78

C.3. Snow use-case

We evaluate the regression capability of THOR on the fractional snow cover task (Tab. S.14). A linear decoder (using TerraTorch's LinearDecoder) trained on frozen THOR features consistently outperforms the fully supervised UNet baseline (RMSE 12.4), DOFA + UPerNet (RMSE 13.8), and THOR-B + UPerNet with patch sizes of 16×16 (RMSE 14.0) and 8×8 (RMSE 12.4). DOFA was selected for comparison due to its wavelength dependent dynamic weight generator that functions as a translation layer for the wavelengths of the Sentinel-3 SLSTR sensor data. Notably, THOR-B with a linear decoder achieves the state-of-the-art RMSE of 9.88, marginally surpassing the UPerNet head (RMSE 9.90). However, THOR-S + UPerNet with 4×4 patch size performed best with an RMSE of 9.69. Most critically, the linear decoder achieves this performance using only 24.6k parameters, compared to the 22.9M parameters required by the UPerNet head. This 1000x reduction in decoder complexity demonstrates that THOR's pre-trained representations are linearly separable and semantically rich, requiring minimal adaptation for downstream physical variable mapping.

C.4. ERA5 Land analysis

To validate the climate-awareness of the frozen encoder, we analyze the performance of the linear probe on the hold-out set against the ground truth ERA5-Land daily statistics variables by sampling random crops with a ground cover of 11520 m and extracting Sentinel-3 OLCI and SLSTR. Fig. S.9 presents scatter plots for these targets, revealing a clear distinction in performance: thermodynamic state variables (e.g., `temperature_2m`, `surface_pressure`)

Table S.10. Extended Pangaea results with 100% training data in mIoU. Bold/underline mark best/second-best per column.

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	FBP	DynEarthNet	CropMap	SN7	AI4Farms	Avg. Rank
CROMA	82.42	<u>67.55</u>	32.32	<u>90.89</u>	51.83	38.29	49.38	59.28	25.65	7.00
DOFA	80.63	59.58	30.02	89.37	43.18	<u>39.29</u>	51.33	61.84	27.07	8.33
GFM-Swin	76.90	64.71	21.24	72.60	67.18	34.09	46.98	60.89	27.19	10.67
Prithvi	<u>83.62</u>	49.98	33.93	90.37	46.81	27.86	43.07	56.54	26.86	12.11
RemoteCLIP	76.59	60.00	18.23	74.26	**69.19**	31.78	52.05	57.76	25.12	12.56
SatlasNet	79.96	55.86	17.51	90.30	50.97	36.31	46.97	61.88	25.13	10.67
Scale-MAE	76.68	57.32	24.55	74.13	<u>67.19</u>	35.11	25.42	**62.96**	21.47	12.44
SpectralGPT	80.47	57.99	35.44	89.07	33.42	37.85	46.95	58.86	26.75	10.67
S12-MoCo	81.58	51.76	34.49	89.26	53.02	35.44	48.58	57.64	25.38	11.00
S12-DINO	81.72	49.37	36.18	88.61	51.15	34.81	48.66	56.47	25.62	12.11
S12-MAE	81.91	49.90	32.03	87.79	51.92	34.08	45.8	57.13	24.69	13.56
S12-Data2Vec	81.91	44.36	34.32	88.15	48.82	35.90	54.03	58.23	24.23	11.89
TerraMindv1-B	82.42	**69.52**	<u>40.51</u>	90.62	59.72	37.87	55.80	60.61	28.12	**3.56**
UNet Baseline	**84.51**	54.79	31.60	**91.42**	60.47	**39.46**	47.57	<u>62.09</u>	**46.34**	<u>5.00</u>
ViT Baseline	81.58	48.19	38.53	87.66	59.32	36.83	44.08	52.57	<u>38.37</u>	11.11
THOR-T	79.34	53.82	38.02	89.35	46.41	33.59	50.39	60.61	26.36	11.11
THOR-S	79.26	52.66	39.54	90.14	47.25	34.84	<u>59.49</u>	60.01	26.91	9.11
THOR-B	79.65	51.48	**40.76**	89.44	47.42	37.57	56.78	59.87	26.29	8.78
THOR-L	79.47	53.73	39.88	89.55	47.62	37.29	**60.75**	59.71	26.75	8.33

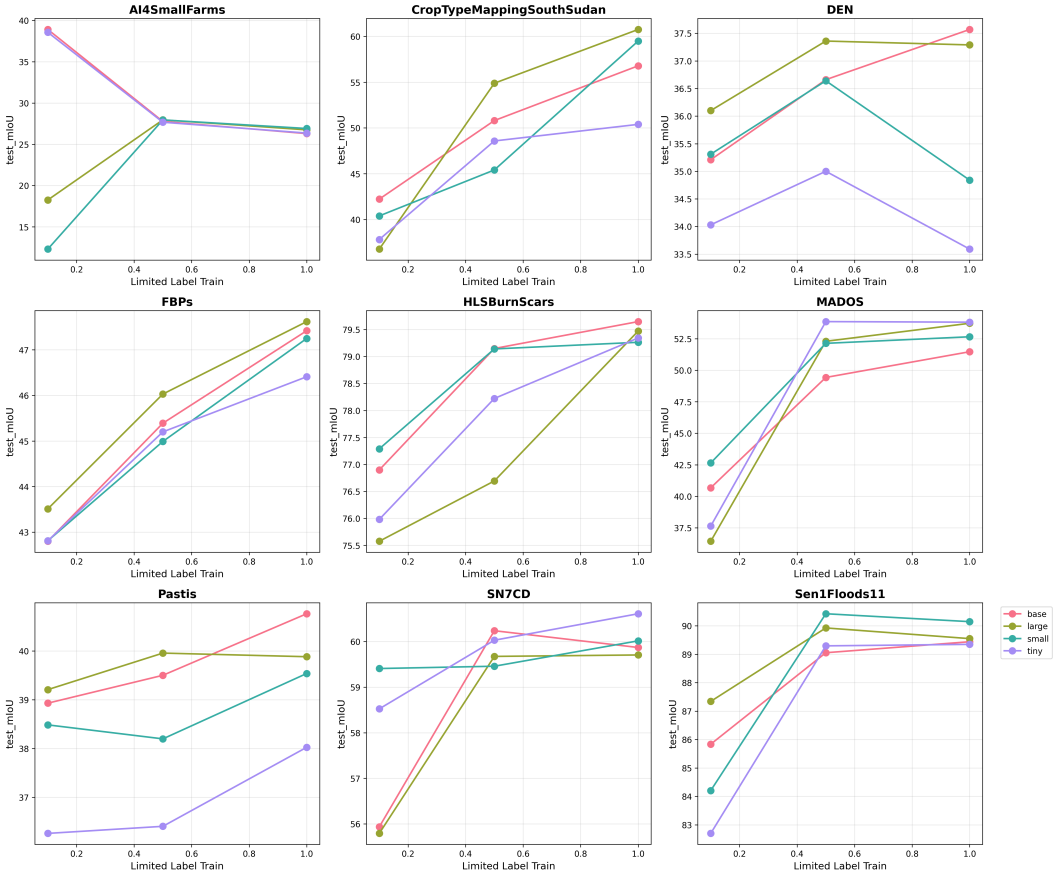
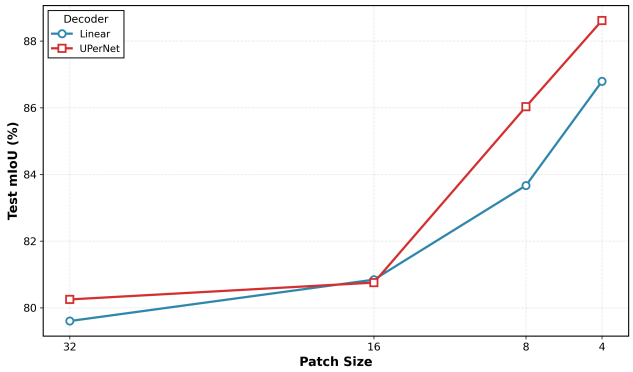
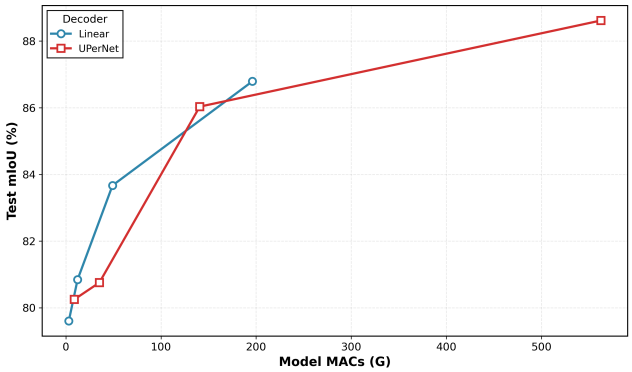


Figure S.5. Per dataset mIoU for all Pangaea benchmarks using 10, 50 and 100% training data for Tiny, Small, Base and Large model. Patch size 6, concat feature aggregation and input image size of 108 pixels.

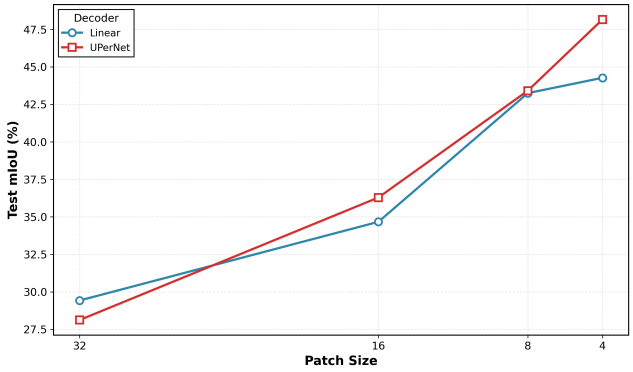
exhibit strong linearity and tight clustering ($R^2 > 0.8$), whereas stochastic, accumulated phenomena (e.g., `snow_depth`, `total_precipitation`) remain challenging to regress from instantaneous optical/SAR snapshots. This trend is quantified in Fig. S.7, which shows low NRMSE and high R^2 for thermal and vegetation indices, contrasting with higher error rates for hydrological variables. However, the structural fidelity of the learned rep-



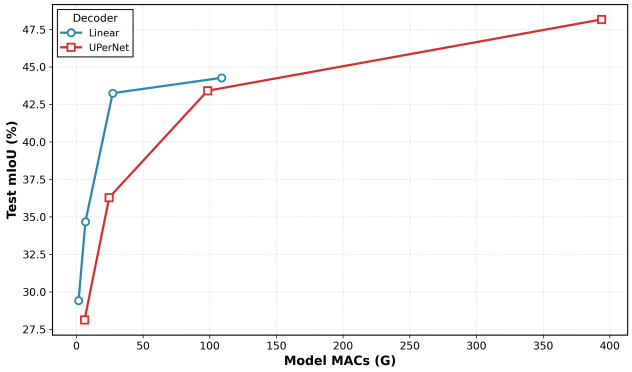
(a) Sen1Floods11 - Patch Size



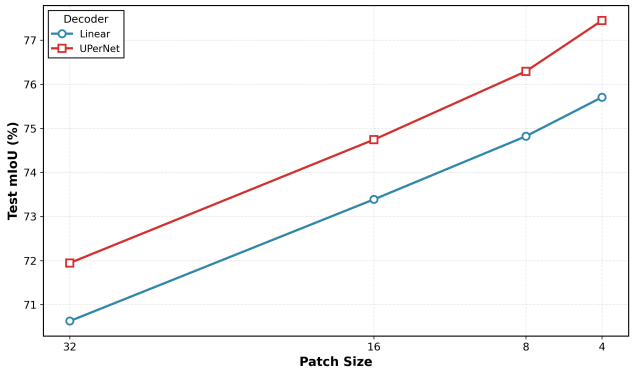
(b) Sen1Floods11 - Model MACs



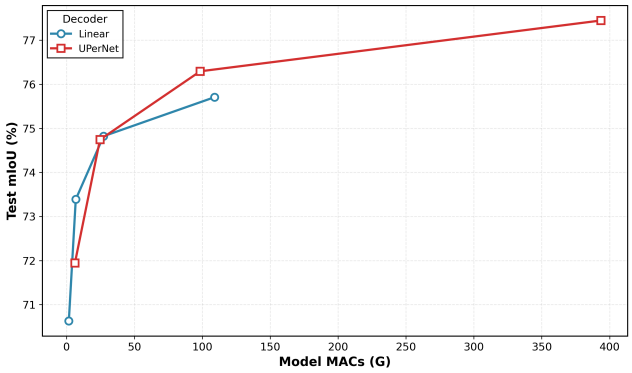
(c) MADOS - Patch Size



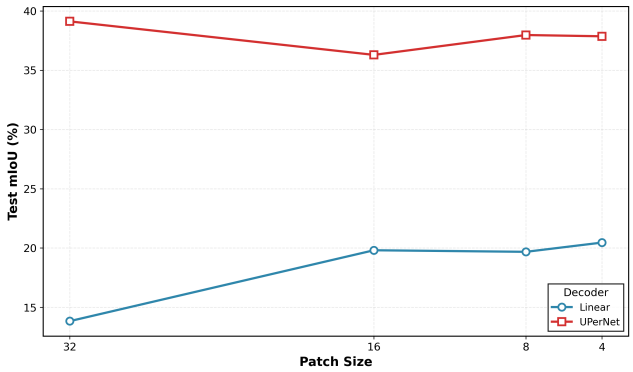
(d) MADOS - Model MACs



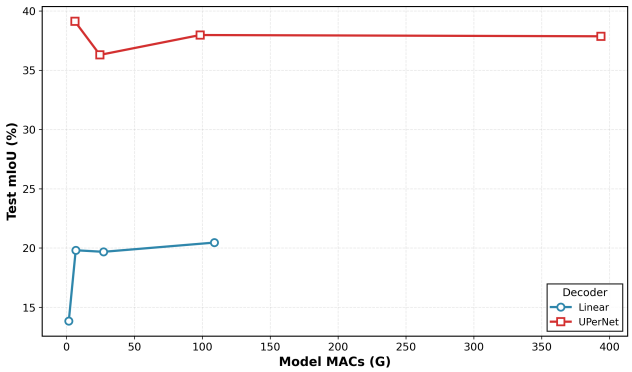
(e) HLS Burns - Patch Size



(f) HLS Burns - Model MACs



(g) AI4Smallfarms - Patch Size



(h) AI4Smallfarms - Model MACs

Figure S.6. Model performance across different datasets. Left column shows test mIoU versus patch size, right column shows the respective test mIoU versus model MACs (G). Using THOR-B frozen encoder and linear decoder $\sim$ 0.2 M parameters (blue circles) and UPerNet decoder $\sim$ 67 $-$ 109M parameters (red squares) are compared across four benchmark datasets. Using a fixed input size of 128, and concatenation of feature maps. All experiments run with 10% training data.

Table S.11. Mean Pangaea test mIoU by output aggregation method and training data, THOR-B model.

Output Aggregation	10%	100%
concat	52.94	58.63
mean	49.90	56.02

Table S.12. Mean Pangaea test mIoU by patch size and training data, THOR-B model.

Patch Size	10%	100%
4	52.94	58.63
6	50.82	54.36
8	52.09	54.69

resentation is confirmed in the correlation matrix of the predicted ERA5-Land values closely mirrors that of the ground truth , demonstrating that THOR successfully captures the physical inter-dependencies between these climatic variables (such as the coupling between soil moisture and temperature) (Fig. S.8). This suggests that the encoder moves beyond visual texture matching to embed the broad climatological context required for downstream climate applications.

Supplementary references

[S1] Hasan Abed Al Kader Hammoud, Tuhin Das, Fabio Pizzati, Philip HS Torr, Adel Bibi, and Bernard Ghanem. On pretraining data diversity for self-supervised learning. In *European Conf. Computer Vision*, pages 54–71. Springer, 2024.

[S2] Olivier Arino, Jose Julio Ramos Perez, Vasileios Kalogirou, Sophie Bontemps, Pierre Defourny, and Eric Van Bogaert. Global Land Cover Map for 2009 (GlobCover 2009), 2012.

[S3] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19530–19540, 2025.

[S4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proc. IEEE/CVF International Conf. Computer Vision*, pages 16772–16782, 2023.

[S5] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *Adv. Neural Inform. Process. Syst.*, 35:197–211, 2022.

[S6] Mark A. Friedl and Damien Sulla-Menashe. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500 m SIN Grid V061 (MCD12Q1). NASA LP DAAC, 2022.

[S7] Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Adv. Neural Informa. Process. Syst.*, 36:5506–5538, 2023.

[S8] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

[S9] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology, 2024. arXiv:2404.10242 [cs].

[S10] 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. *Adv. Neural Inform. Process. Syst.*, 36:51080–51093, 2023.

[S11] Jeremy Lai, Faruk Ahmed, Supriya Vijay, Tiam Jaroensri, Jessica Loo, Saurabh Vyawahare, Saloni Agarwal, Fayaz Jamil, Yossi Matias, Greg S Corrado, et al. Domain-specific optimization and diverse evaluation of self-supervised models for histopathology. *arXiv preprint arXiv:2310.13259*, 2023.

[S12] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

[S13] Sari Metsämäki, Jouni Pulliainen, Miia Salminen, Kari Luojus, Andreas Wiesmann, Rune Solberg, Kristin Böttcher, Mwaba Hiltunen, and Elisabeth Ripper. Introduction to GlobSnow Snow Extent products with considerations for accuracy assessment. *Remote Sensing of Environment*, 156:96–108, 2015.

[S14] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European*

Table S.13. GEO-Bench Image classification test performance (%) via kNN.

		m-EuroSat Training %, Top-1 Acc. ↑				**m-BigEarthNet** Training %, F1 Score ↑				**m-So2Sat** Training %, Top-1 Acc. ↑				**m-Brick-Kiln** Training %, Top-1 Acc. ↑			
Method	Arch.	100%	20%	5%	1%	100%	20%	5%	1%	100%	20%	5%	1%	100%	20%	5%	1%
SatMAE [S5]	ViT-Base	84.1	73.3	50.1	34.8	50.6	42.5	35.7	29.0	36.0	32.9	29.7	23.1	86.1	81.9	80.3	73.5
SatMAE	ViT-Large	84.3	74.7	53.1	46.4	50.8	42.9	35.6	27.7	36.6	34.3	31.0	24.4	87.9	84.0	80.4	74.7
SatMAE++ [S15]	ViT-Large	82.7	75.9	51.1	48.5	50.8	42.8	36.7	31.6	34.7	32.7	29.9	23.4	89.6	87.1	82.8	76.7
CROMA [S7]	ViT-Base	85.6	79.4	66.2	51.3	58.8	55.3	49.3	44.7	48.8	48.0	43.9	33.8	92.6	90.6	87.7	85.1
CROMA	ViT-Large	86.3	78.1	59.9	49.0	56.6	50.6	44.1	38.0	47.6	45.0	43.2	33.7	91.0	86.7	82.9	80.2
SoftCon [S21]	ViT-Small	89.8	83.4	55.9	27.2	**64.7**	58.7	52.6	43.3	51.1	49.9	43.3	31.4	89.2	86.9	80.5	77.8
SoftCon	ViT-Base	90.3	82.1	54.2	19.8	63.7	57.5	52.0	42.5	51.0	49.7	45.3	35.4	90.0	86.1	80.6	74.5
DOFA-v1 [S22]	ViT-Base	82.8	72.1	60.9	49.6	49.4	43.6	37.2	29.9	41.4	40.7	37.5	29.4	88.3	86.2	82.0	78.3
DOFA-v1	ViT-Large	83.6	72.1	53.5	41.7	49.9	41.6	35.3	27.6	45.4	40.6	35.6	31.8	86.8	85.2	84.8	80.6
Satlas [S4]	Swin-Tiny	81.7	70.3	48.3	35.8	51.9	44.8	37.8	29.6	36.6	30.7	29.6	27.1	88.2	85.2	82.4	73.0
Satlas	Swin-Base	81.5	69.1	42.1	10.0	47.0	41.1	35.0	25.8	35.8	33.4	29.6	30.4	80.0	78.3	76.9	73.3
MMEarth [S14]	CNN-atto	81.7	73.5	60.3	30.0	58.3	52.2	46.5	39.6	39.8	38.8	36.8	25.1	89.4	85.4	84.1	79.7
DeCUR [S20]	ViT-Small	89.0	85.3	**72.3**	46.6	63.8	**59.2**	**55.4**	**49.6**	45.8	43.1	38.5	30.9	83.7	81.7	77.9	74.2
Prithvi 2.0 [S18]	ViT-Large	80.2	69.4	54.1	48.0	49.4	42.9	35.5	28.8	29.5	31.2	29.6	26.1	87.9	86.8	83.3	80.6
AnySat [S3]	ViT-Base	82.2	73.7	62.5	47.1	54.9	47.2	40.7	33.7	39.8	34.9	32.0	29.0	85.3	81.7	78.0	72.0
Galileo [S19]	ViT-Nano	89.7	82.4	56.6	41.7	53.8	46.3	41.5	33.9	50.1	50.3	47.5	37.4	86.7	82.2	83.2	79.7
Galileo	ViT-Tiny	90.1	83.9	59.5	41.3	55.5	48.2	41.6	34.4	49.7	50.5	44.2	36.2	86.9	83.7	83.8	77.3
Galileo	ViT-Base	**93.0**	**88.5**	71.3	**56.6**	59.0	51.5	45.4	36.5	**54.8**	**53.8**	**51.1**	**43.2**	90.7	86.9	85.8	78.0
THOR	ViT-Tiny	84.4	74.9	65.4	52.8	56.9	48.9	42.3	37.1	36.4	33.4	33.5	28.0	92.9	91.6	90.1	85.2
THOR	ViT-Small	84.3	78.9	66.5	49.3	58.4	49.7	43.0	37.1	37.2	35.1	33.5	29.4	**95.5**	**94.1**	**93.0**	**91.5**
THOR	ViT-Base	84.9	76.3	66.2	52.0	58.3	49.3	41.8	37.0	37.9	35.3	31.3	26.2	94.6	93.6	90.6	88.2
THOR	ViT-Large	85.8	76.5	64.3	51.5	58.3	50.1	42.6	35.6	41.0	36.6	37.2	27.6	94.0	93.2	91.3	84.6

Table S.14. RMSE snow cover fraction. Image size 128×128 and concatenated the tokens of the 500 m and 1000 m bands. For DOFA we sampled images of size 112×112 that was resized to 224×224 to meet DOFA's input image size requirement.

Decoder	Encoder	Patch size	No. dec. param.	Tot. no. param.	RMSE
UNet				24.4M	12.4
UPerNet	DOFA-B	16x16	12.1M	0.1G	13.8
UPerNet	THOR-B	16x16	22.9M	0.1G	14.0
		8x8			12.4
		4x4			9.90
	THOR-T	4x4	8.6M	16.2M	10.5
	THOR-S	4x4	12.1M	37.9M	**9.69**
	THOR-L	4x4	33.1M	0.3G	12.2
Linear decoder	THOR-B	4x4	24.6k	94.2M	9.88
	THOR-T	4x4	6.1k	7.6M	11.5
	THOR-S	4x4	12.3k	25.8M	10.9
	THOR-L	4x4	32.8k	0.3G	10.3

Conf. Computer Vision, pages 164–182. Springer, 2024.

[S15] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proc. IEEE/CVF Conf. Computer Vision Pattern Recog.*, pages 27811–27819, 2024.

[S16] Chris Oliver and Shaun Quegan. *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.

[S17] A Ordonez, D Wade, C Ravaut, and AU Waldeland. Towards a foundation model for seismic interpretation. In *85th EAGE Ann. Conf. & Exhibition (including the Workshop Programme)*, pages 1–5. European Association of Geoscientists & Engineers, 2024.

[S18] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications.

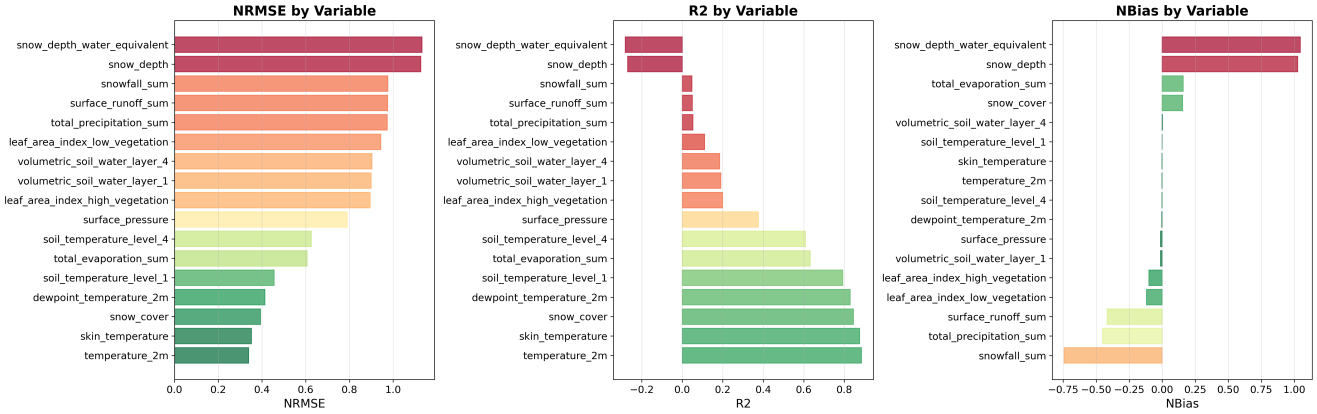


Figure S.7. Comparison of NRMSE, R^2, and normalized bias for 17 ERA5 variables predicted from satellite embeddings. Color scale ranges from green (good performance) to red (poor performance), with bias colored to highlight deviations from zero.

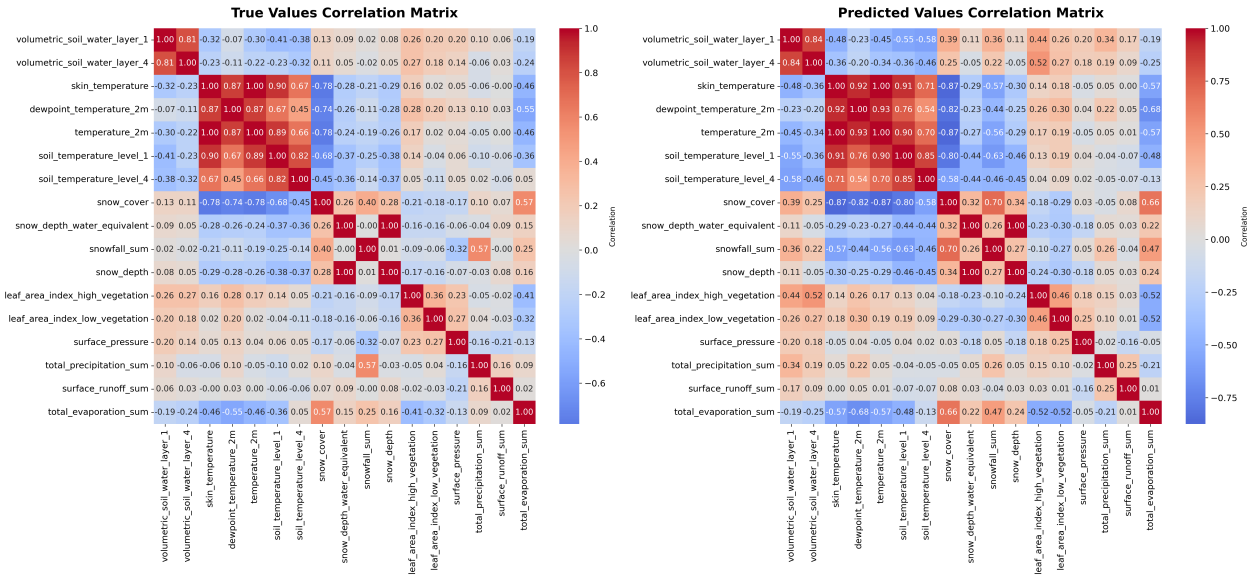


Figure S.8. Comparison of inter-variable correlations for ground truth (left) and model-predicted (right) ERA5 variables.

arXiv preprint arXiv:2412.02732, 2024.

[S19] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *arXiv preprint arXiv:2502.09356*, 2025.

[S20] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conf.Computer Vision*, pages 286–303. Springer, 2024.

[S21] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Multilabel-guided soft contrastive learning for efficient earth observation pretraining. *IEEE Trans. Geosci. Remote Sensing*, 62:1–16, 2024.

[S22] 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, 2024. arXiv:2403.15356 [cs].

[S23] Daniele Zanaga, Ruben Van De Kerchove, Wanda De Keersmaecker, Niels Souverijns, Carsten Brockmann, Ralf Quast, Jan Wevers, Alex Grosu, Audrey Paccini, Sylvain Vergnaud, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Ivelina Georgieva, Myroslava Lesiv, Sarah Carter, Martin Herold, Linlin Li, Nandin-Erdene Tsendbazar, Fabrizio Ramoino,

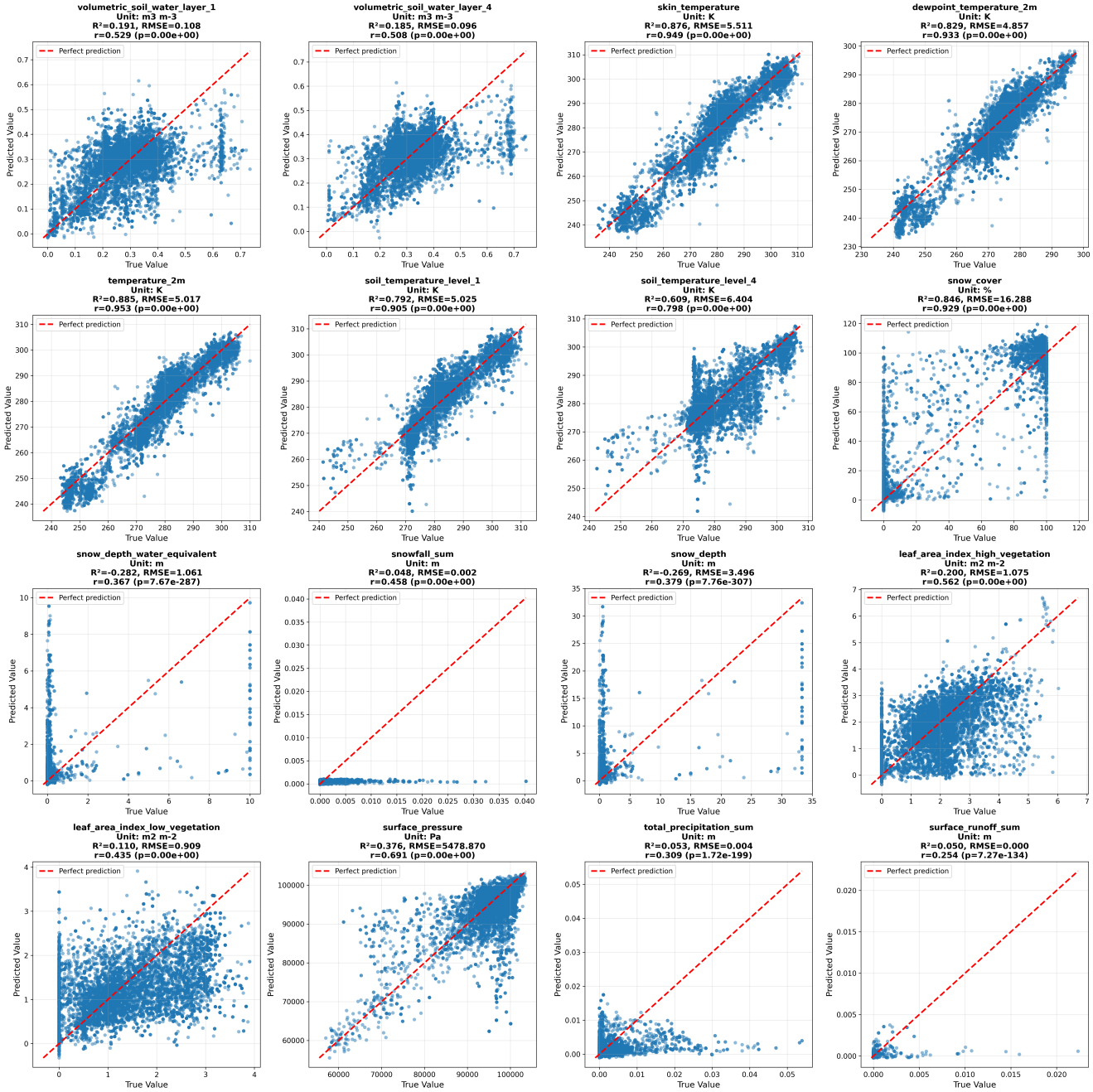


Figure S.9. Model predictions plotted against ground truth observations for 17 ERA5 variables. Red dashed lines indicate perfect predictions.

and Olivier Arino. Esa worldcover 10 m 2020 v100, 2021.