Evaluating the Diversity and Representativeness of Large-Scale and General-Purpose AI4EO Datasets

Syed Roshaan Ali Shah[1], Muhammad Akhtar Munir[2], Muhammad Sohail Danish[2], Syed Waqas Zamir[3], Fahad Shahbaz Khan[2], Salman Khan[2,4]

[1]University College London [2]Mohamed Bin Zayed University of AI [3]Microsoft AI for Good Research Lab [4]Australian National University

A Multiaxial Framework

- Rapid growth in AI4EO has produced many datasets, yet their general-purpose utility often remains constrained, partly due to limited diversity and representativeness.
- We assess 15 prominent AI4EO datasets across 6 key axes: geography, land cover, climate, biomes, human activity, and EO themes.

The goal is to identify structural and semantic diversity gaps, expose imbalances in coverage across ecological and human dimensions, and promote proportional, ecologically valid sampling strategies to support generalizable, inclusive, and context-aware future AI4EO dataset curation.

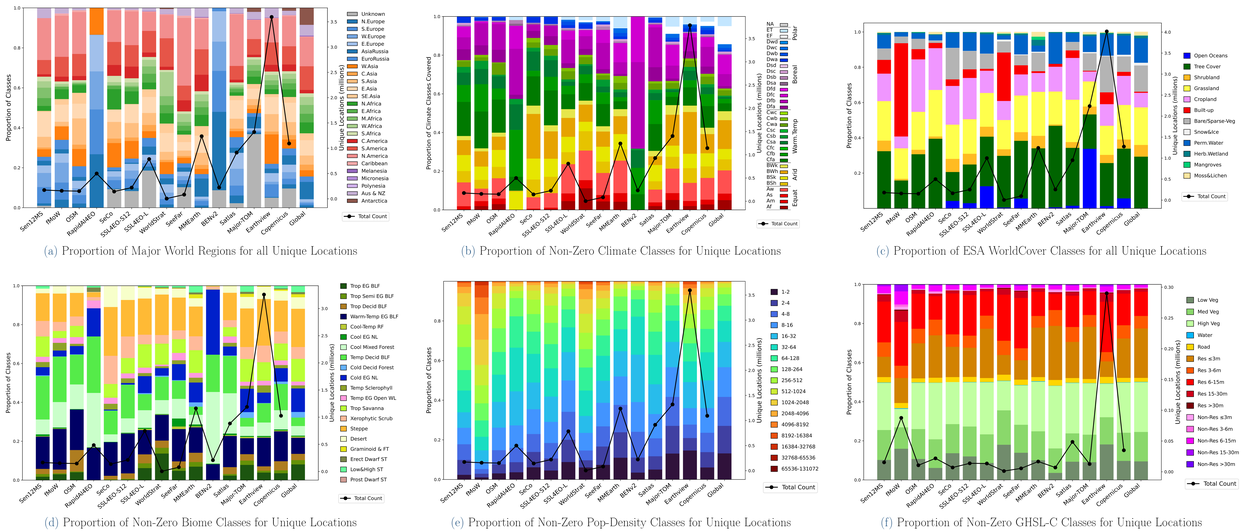
- Our data and code are available at: `https://github.com/mbzuai-oryx/NestEO`

List of the Evaluated Datasets

No	Dataset	Year	Modality	Rez (m/pix)	Grid/Patch size	Count(≈)	Locations	Timestamps	Coverage
1	Sen12MS	2019	S2, S1	10	256 × 256	180662	180662	Single	Global
2	fMoW	2018	RGB/full	≈0.5	variable	1047691	45664	Variable	Global
		2022	S2	10	variable	882779	113382	Variable	Global
3	OpenSentinelMap	2022	S2	10	192 × 192	137045	137045	2	Global
4	RapidAI4EO	2021	S2 L2A	10	60 × 60	500,000	500,000	Dense	Europe
			Planet	3	200 × 200				
5	SeCo	2021	S2 L2A	10	265 × 265	1M	200,000	5	Global
6	SSL4EO-S12 v1.1	2023	S2 L1C, L2A	10	264 × 264	≈1M	246,144	4	Global
			S1 GRD	10					
7	SSL4EO-L	2023	TOA & SR	30	264 × 264	5M	1M	4	Global
8	WorldStrat	2022	SPOT 6/7	1.5	1054 × 1054	62,848	3,928	1	Global
			S2 L1C, L2A	10	158 × 158			16	
9	See-Far	2024	NewSat IV	1	384 × 384	33611	100	736	Global
			Spot 6/7	1.5			29347	≈1	
			S2 L2A	10			32800	≈1	
			Landsat	30			33457	≈1	
10	MMEarth	2024	S2 L2A, S1 GRD	10	128 × 128	1,243,233	1,243,233	Single	Global
11	BigEarthNetv2.0	2024	S2 L2A, S1 GRD	10	120 × 120	549,488	549,488	Single	Europe
12	Satlas	2023	NAIP	1	8192 × 8192	54,987,761	11,868,302	Variable	US
		2023	S2 L1C	10	512 × 512	-	950,490	8–12	Global
		2024	Landsat	10	512 × 512	-	≈950,490	8–12	Global
		2024	S1 GRD	10	512 × 512	-	≈950,490	8–12	Global
13	Major-TOM	2024	S2 L1C, L2A	10	1068 × 1068	2,245,886	2,238,689	≈1	Global
			S1 RTC	10	1068 × 1068	1,469,955	1,458,481	≈1	Global
14	Earthview	2025	Satellogic	1	384 × 384	≈7M	2,967,464	1–5	Global
			S2, S1	Native	variable	≈10M, ≈5M	1,049,409	3–10	Global
			NEON-RGB	0.1	640 × 640	1386	462	3	US
			NEON-HS/Lidar	1	64 × 64	1386	462	3	US
15	Copernicus-Pretrain	2025	S1 GRD, S2 TOA	10	264 × 264	4227387	1067267	≈4	Global
			Sentinel-3 OLCI	300	96 × 96	2189561	281375	≈8	Global
			S-5P CO,NO2,SO2,O3	1000	28 × 28	≈1.3 - 2.5M	≈300k	1–12	Global

Likert-Scale Style Scoring Criteria

Criteria	1	2	3	4	5
Geographical Regions	Localized	Constrained	Clustered	Global	Dense Global
Landcovers	1–3 Classes	Few Classes	Major Classes	All Classes	Sensibly Balanced & Diverse
Biomes	1–3 Classes	Few Classes	Major Classes	All Classes	Sensibly Balanced & Diverse
Climates	1–5 Classes	Few Classes	Major Classes	All Classes	Sensibly Balanced & Diverse
Human Activity Zones	Unsampled	Missing Classes	Major Classes	All Classes	Sensibly Balanced & Diverse
EO themes	One Theme	Few Themes	Moderate	Broad Scope	Thematically Complete



(a) Proportion of Major World Regions for all Unique Locations (b) Proportion of Non-Zero Climate Classes for Unique Locations (c) Proportion of ESA WorldCover Classes for all Unique Locations (d) Proportion of Non-Zero Biome Classes for Unique Locations (e) Proportion of Non-Zero Pop-Density Classes for Unique Locations (f) Proportion of Non-Zero GHSL-C Classes for Unique Locations

Figure 1. Diversity analyses across regions, climates, landcovers, biomes, population and urban characteristics. OSM= OpenSentinelMap & BENv2= BigEarthNetv2.0. For Biomes; NL= Needleleaf, BLF= Broadleaf Forest, RF= Rainforest, WL= Woodland, EG= Evergreen, Decid= Deciduous, Temp= Temperate, Trop= Tropical, Prost= Prostrate, ST= Shrub Tundra, FT= Forb Tundra. Disparities are evident in coverage proportions. Most datasets over-sample temperate forests and urban centers; few reflect tropical deserts, tundras, or rural low-density zones. Excluding Major-TOM's large coverage of Open Oceans, MMEarth, Earthview, Major-TOM, and Copernicus approach global proportionality across more axes.

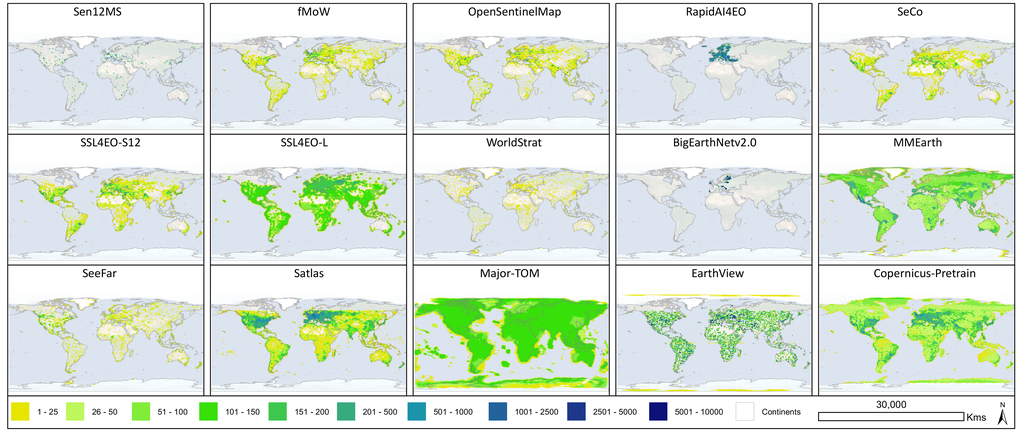


Figure 2. Spatial sampling densities of existing Earth Observation (EO) datasets projected onto Level 3 H3 hexagonal discrete global grids. The figure illustrates the geographic distribution biases, regional imbalances, and density variations across datasets, highlighting significant oversampling in certain areas (e.g., Europe, North America) and undersampling in others (e.g., central Africa, parts of Oceania). Strong urban-centric or Eurocentric biases are observed in Satlas, RapidAI4EO, and fMoW.

Composite Scores Across All Criteria

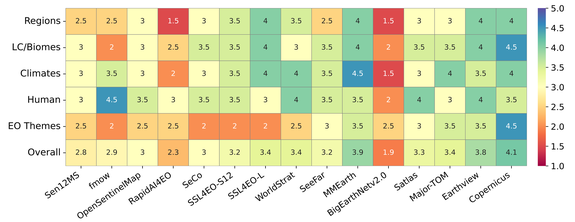


Figure 3. Criteria Scorings for the studied datasets: No dataset scores uniformly very high across all criteria. Overall, Copernicus-Pretrain, Earthview, and MMEarth emerge as the better-balanced and globally representative datasets. Other datasets exhibit domain-specific strengths but show gaps in other diversity metrics.

Key Observations

- Most datasets disproportionately sample temperate forests, with urban cores variably represented
- Only 3 out of 15 datasets exhibit reasonable and balanced proportionality across coverage dimensions
- Major-TOM assigns over 30% samples to ocean tiles, diluting land diversity metrics
- Under-sampling of Africa and Oceania persists across the majority of datasets

Key Design Lessons

- Size ≠ Diversity: true representativeness stems from proportional sampling and thematic breadth, not just volume
- Uniform global sampling can dilute signals from minority classes, weakening semantic learning & contrastive signals in SSL models
- Oversampling rare but semantically rich classes can improve generalization
- Algorithms benefit from structured and semantically rich data & Semantic + Sampling structure of the dataset matters

Redesigning Sampling: From Coverage to Compositional Semantics

- Use proportion-aware stratified sampling over ecological and anthropogenic axes
- Include and oversample semantically rare but important classes
- Include multi-modal, multi-platform and multi-scale integrations
- Use of hierarchical geospatial grids for representative, aligned & scalable curation

Final Takeaway

Uniform coverage is not enough. Curated proportionality, semantic coherence, and class structure are the missing links in today's AI4EO datasets

Nested Tiling + Proportional Sampling + Thematically balanced curation –> NestEO