SHRUG-FM: Reliability-Aware Foundation Models for Earth Observation

Supplementary Material

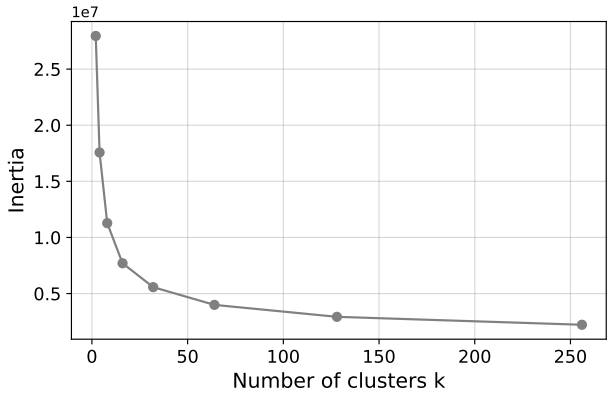

Figure 3. **Prototype selection for embedding familiarity.** k-means clustering of pretraining embeddings. The elbow criterion indicates stable performance around $k = 64$.

7. Appendix: Additional Results

We provide a set of ablations regarding the importance and usage of the different reliability signals in Tab. 3 and Tab. 4.

8. Appendix: Implementation Details for Selective Classifier

In the very first step, we remove highly correlated physical features. Next, feature subsets are selected using RFECV with decision trees over all candidate signals $(\mathcal{U}_I, \mathcal{U}_E, \mathcal{U}_T)$.

For model selection, we tune decision tree hyperparameters using cross-validation. Fire and landslide tasks use repeated, stratified k-fold validation, while flood detection uses standard k-fold validation. The maximum tree depth is limited to three in all experiments.

Final models are trained on the combined training and validation splits. Reported performance reflects evaluation on the held-out test set.

9. Appendix: Feature Selection and Stability Analysis

To ensure the robustness of the SHRUG-FM framework, we repeat the entire pipeline—including RFE-based feature selection and tree tuning—across 30 independent runs with varying random seeds. We observe minimal sensitivity to the seed, indicating that our framework successfully identifies stable, physically-grounded failure patterns.

	Classifier ROC AUC↑			Risk–Coverage AUC↓		
	Fire	Flood	Landslides	Fire	Flood	Landslides
All	0.791 ± 0.0635	0.725 ± 0.00378	0.654 ± 0.0263	0.16 ± 0.0435	0.114 ± 0.00124	0.392 ± 0.0329
No Input Feature Signals	**0.84 ± 0.00944**	**0.739 ± 0.00106**	**0.661 ± 0.0382**	**0.13 ± 0.0105**	0.104 ± 0.000206	**0.381 ± 0.0385**
No Embedding OOD Signals	0.77 ± 0.0224	0.724 ± 0.00365	0.622 ± 0.0373	0.173 ± 0.0258	0.114 ± 0.0012	0.423 ± 0.0823
No UQ (MI, AE) Signals	0.706 ± 0.0242	0.524 ± 0.024	0.548 ± 0.0248	0.227 ± 0.0282	0.207 ± 0.00605	0.499 ± 0.00608
UQ only	0.8 ± 0.00788	0.739 ± 0.000477	0.633 ± 0.0104	0.138 ± 0.0102	**0.104 ± 0.000229**	0.412 ± 0.039
Dist only	0.739 ± 0.0465	0.495 ± 0.00921	0.562 ± 0.0181	0.198 ± 0.0382	0.21 ± 0.00418	0.494 ± 0.00732
Input only	0.473 ± 0.0556	0.533 ± 0.0056	0.417 ± 0.0529	0.351 ± 0.0404	0.205 ± 0.00193	0.604 ± 0.0495

Table 3. Ablation of reliability signals. Selective prediction performance using input-only, embedding-only, task-only, and combined signals. For each scenario, we withheld a subset of the signals for the whole selective prediction training pipeline.

feature	Fire	Flood	Landslide
`Average_Entropy (0.05)`	—	100.0%	16.7%
`Average_Entropy (0.3)`	90.0%	—	—
`Mutual Information (0.05)`	—	100.0%	100.0%
`Mutual Information (0.1)`	100.0%	—	—
`density`	53.3%	100.0%	—
`ncdd`	30.0%	—	70.0%
`normalized_distance`	—	—	100.0%
`ria_ha_ssu`	—	—	3.3%

Table 4. Signals used by the selective classifier. Most informative reliability indicators were identified via recursive feature elimination and feature importance of the resulting tree. Decimals after the uncertainty scores denote the threshold on the predicted probability, marking the region of interest over which the measure is averaged for the image-level uncertainty.

Table 5. Full set of input features for the selective classifier g.

Category	Feature	Description
HydroATLAS	`inu_pc_smn`	Min. annual inundation extent (% cover).
	`inu_pc_smx`	Max. annual inundation extent (% cover).
	`ria_ha_ssu`	Total river area (hectares).
	`slp_dg_sav`	Average terrain slope (degrees × 10).
	`snw_pc_syr`	Avg. annual snow cover extent (% cover).
	`snw_pc_smx`	Max. annual snow cover extent (% cover).
	`glc_cl_smj`	Land cover classes (spatial majority).
	`wet_pc_sg1`	Wetland extent (grouping 1, % cover).
	`wet_pc_sg2`	Wetland extent (grouping 2, % cover).
	`for_pc_sse`	Average forest cover extent (% cover).
	`crp_pc_sse`	Average cropland extent (% cover).
	`pst_pc_sse`	Average pasture extent (% cover).
	`ire_pc_sse`	Average irrigated area extent (% cover).
	`urb_pc_sse`	Average urban extent (% cover).
	`hft_ix_s09`	Human Footprint index (year 2009).
DEM	`elev_mean`	Average elevation (meters a.s.l.).
	`elev_min`	Minimum elevation (meters a.s.l.).
	`elev_max`	Maximum elevation (meters a.s.l.).
Spatial	`density`	Spatial density of pre-training data.
Embedding	`NCDD`	Centroid Distance Deficit OOD score.
	`norm_dist`	Distance to closest pre-training cluster.
Uncertainty	`avg_entropy`	Image-level average aleatoric uncertainty.
	`mutual_info`	Image-level average epistemic uncertainty.

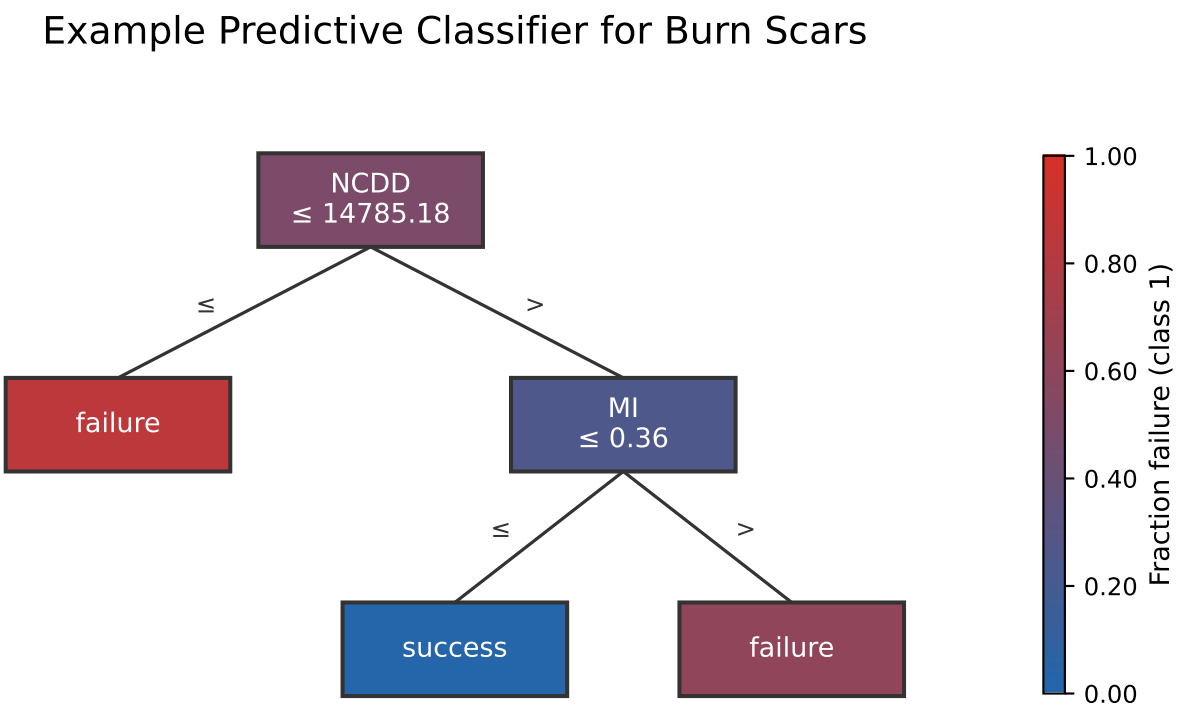

Figure 4. **Example selective classifier.** Shallow decision tree learned from reliability signals on the burn scars task. The model uses NCDD as a first signal. When it lies below a threshold which indicates a stronger out of distribution signal, it classifies as failure. Above the threshold, mutual information corresponding to epistemic uncertainty is used for another branching-off with high uncertainty indicating failure and low uncertainty success. The structure reflects intuitive failure modes, combining complementary signals from the scores.

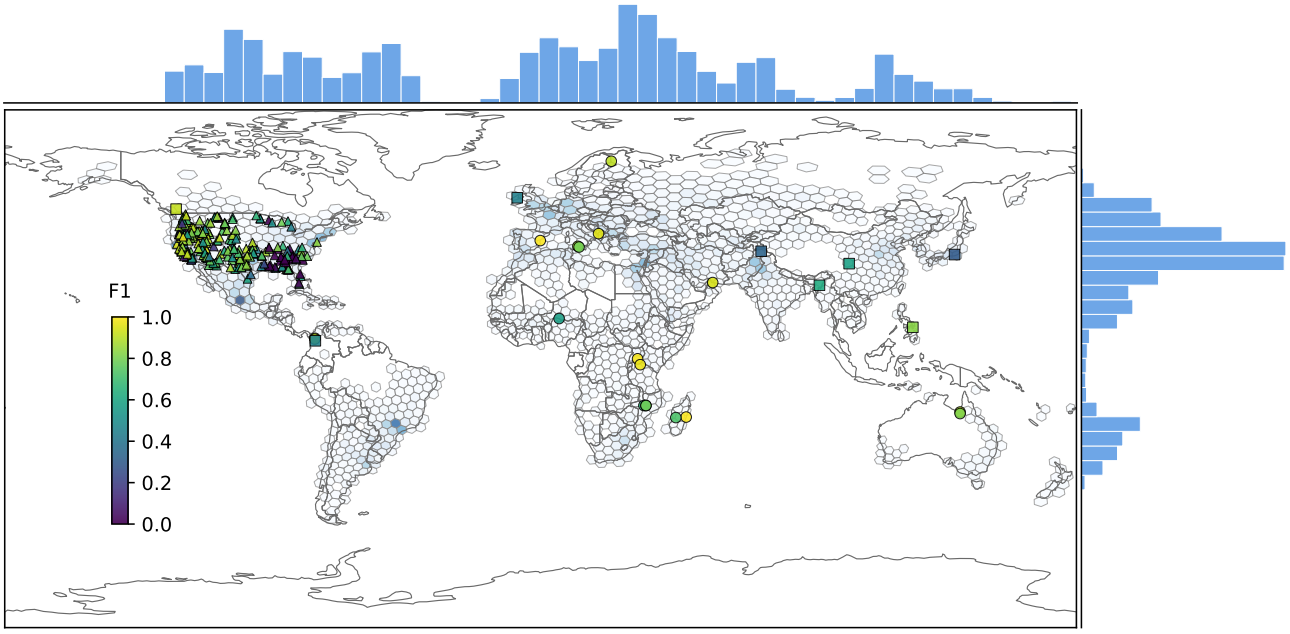

Figure 5. **Data distributions and performance.** F1 performance on the test set of the selected downstream tasks, overlaid on a hexagonal spatial distribution of SSL4EO-S12 across the globe. Shapes identify the task: △ - burn scars, ◯ - floods and □ - landslides.