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

Kai-Hendrik Cohrs[*1] **Zuzanna Osika**[*2] **Maria Gonzalez-Calabuig**[*1] **Vishal Nedungadi**[*3]
Ruben Cartuyvels[4] **Steffen Knoblauch**[5] **Joppe Massant**[†6]
Shruti Nath[†7] **Patrick Ebel**[†4] **Vasileios Sitokonstantinou**[†31]

[1]Universitat de València, [2]Delft University of Technology, [3]Wageningen University & Research,
[4]European Space Agency, [5]Heidelberg University, [6]Ghent University, [7] University of Oxford
{kai.cohrs, maria.gonzalez-calabuig}@uv.es
z.osika@tudelft.nl
{vishal.nedungadi, vassilis.sitokonstantinou}@wur.nl
{ruben.cartuyvels, patrick.ebel}@esa.int
steffen.knoblauch@uni-heidelberg.de
joppe.massant@ugent.be
shruti.nath@physics.ox.ac.uk

Abstract

Geospatial foundation models for Earth observation often fail to perform reliably in environments underrepresented during pretraining. We introduce SHRUG-FM, a framework for reliability-aware prediction that integrates three complementary signals: out-of-distribution (OOD) detection in the input space, OOD detection in the embedding space and task-specific predictive uncertainty. Applied to burn scar segmentation, SHRUG-FM shows that OOD scores correlate with lower performance in specific environmental conditions, while uncertainty-based flags help discard many poorly performing predictions. Linking these flags to land cover attributes from HydroATLAS shows that failures are not random but concentrated in certain geographies, such as low-elevation zones and large river areas, likely due to underrepresentation in pretraining data. SHRUG-FM provides a pathway toward safer and more interpretable deployment of GFMs in climate-sensitive applications, helping bridge the gap between benchmark performance and real-world reliability.

1 Introduction

Following the success of foundation models in natural language processing and computer vision, geospatial foundation models (GFMs) for Earth observation (EO) are gaining traction [1]. Trained on large-scale datasets gathered by satellites such as Sentinel and Landsat, these models aim to learn transferable representations of the Earth's surface. Recent examples like Clay [2], SSL4EO-S12 [3], Prithvi [4], EarthPT [5] and Scale-MAE [6] capture detailed spatial and spectral patterns across diverse regions and timescales. These general-purpose representations are provided as precomputed embeddings or pre-trained models and can be applied to downstream tasks, out-of-the-box or with minimal finetuning to capture more application-specific features.

However, ongoing research questions the reliability of foundation models for EO in real-world scenarios [7, 8]. These models can fail under spatial extremes, such as underrepresented geographies (e.g., deserts, polar regions, high latitudes etc.) and temporal extremes, such as atypical seasons,

[*] Contributed equally
[†] Supervised

extreme weather events or long-term environmental shifts like droughts, due to gaps in geographic or seasonal coverage of the data used to pre-train the models [9, 7]. Most models lack built-in mechanisms for detecting out-of-distribution (OOD) inputs or for quantifying uncertainty, often producing overconfident predictions. Recent benchmarks such as REOBench [10], GeoBench [11] and PANGAEA [8] evaluate GFMs under conditions that better reflect real-world deployment, including shifts in geography, seasonality and sensor type. The results show that GFMs are regularly outperformed by simple supervised baselines, highlighting the need for systematic reliability assessments of their predictions, including OOD detection and uncertainty estimation.

We present an initial version of SHRUG-FM (Systematic Handling of Real-world Uncertainty for Geospatial Foundation Models), a framework for reliability-aware prediction in GFMs for EO (Figure 1). It integrates three complementary signals: (1) OOD detection in the input space, (2) OOD detection in the embedding space and (3) task-specific predictive uncertainty. The first two identify samples that deviate from the training distribution, either in raw inputs or embeddings, while the third captures uncertainty in the downstream model's predictions. Together, these signals provide a comprehensive reliability assessment, supporting deployment decisions and guiding future pretraining data strategies. We demonstrate SHRUG-FM on burn scar segmentation (ExEBench [7]) using models trained on the SSL4EO-S12 dataset [9], which offers global coverage and multiple pretrained encoders of varying sizes and training strategies, enabling systematic comparisons. We incorporate geospatial context via HydroATLAS [12], overlaying predictions and uncertainties onto layers such as elevation or land cover. This enables users to identify unreliable predictions and link them to specific environmental features, revealing interpretable gaps in the foundation model's pretraining. Such diagnostics are critical for climate-sensitive applications, where overconfident or inaccurate predictions can misguide interventions. While demonstrated on burn scars, SHRUG-FM is designed to extend to other GFMs and tasks and we will evaluate its broader utility in future work.

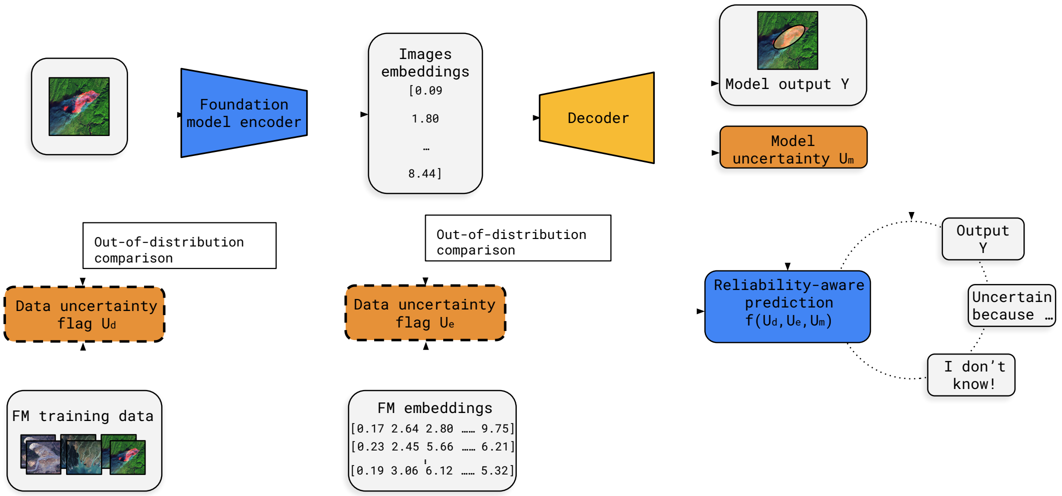

Figure 1: The SHRUG-FM framework. It computes three complementary signals: OOD detection in the input, OOD detection in the embeddings and task-specific predictive uncertainty. These are combined to enable reliability-aware predictions, flagging or abstaining from low-confidence outputs.

2 Data and Methods

Our work builds on SSL4EO-S12 [9], a large-scale global Sentinel-2 dataset (Appendix 5.1), and pretrained models derived from it. These models use different self-supervised learning strategies, contrastive learning (MoCo), self-distillation (DINO) and masked autoencoding (MAE), with a ViT-S/16 backbone. All models are frozen and used as feature extractors for downstream segmentation tasks. We apply SHRUG-FM to burn scar segmentation (ExEBench Burn Scars [7]). Wildfires are increasing in frequency and intensity under climate change [13] and accurate mapping of burn scars is essential for understanding local fire regimes, monitoring ecosystem recovery and supporting risk management [14, 15]. The burn scars dataset captures high-impact events that are often un- [redacted] -world challenges such as cloud cover, class imbalance and high variability in event size and year, making them ideal for studying reliability under distribution shift. For uncertainty estimation and evaluation, we train a decoder on the downstream dataset. For the OOD detection components in the raw input and embedding spaces, we match the spectral bands between SSL4EO-S12 and the downstream data to ensure comparability with the pretraining distribution (Appendix 5.1.2). To interpret the flags, we examine their spatial distribution and link them with HydroATLAS attributes [12] (Appendix Table 1, Fig. 5). This reveals which features are most associated with high uncertainty or distributional mismatch, providing insights into what the GFM may have underrepresented during pretraining.

Out-of-Distribution Detection. To detect inputs that deviate from the pretraining distribution, we analyze their position in both the raw input space and the embedding space of the foundation model. We apply k-means clustering on the SSL4EO-S12 dataset in both spaces, selecting the number of clusters via the heuristic elbow method (k=15). For each test sample, we compute its Euclidean distance to the nearest cluster centroid. We analyze the densities of these distances (Appendix Fig. 3) and observe that downstream samples are generally further away from the pretraining cluster centers. We also compute the Nearest Centroid Distance Deficit (NCDD) [16], which compares the distance to the nearest centroid with the summed distances to all others (Appendix 5.2.1 for more details). These two metrics provide complementary signals: the distance measures how far a sample lies from the nearest known group, while NCDD quantifies how confidently it belongs to that group versus being ambiguous between others.

Uncertainty Quantification. To effectively measure a model's uncertainty, we must distinguish between aleatoric (due to inherent data variability) and epistemic (due to a lack of knowledge) uncertainty [17]. While standard models may provide probability outputs, which can be interpreted as aleatoric uncertainty, these are often overconfident, especially for OOD data [18]. We improve upon these by training an ensemble of decoders (Appendix 5.3.2) on bootstrapped data [19] as well as a model with dropout uncertainty estimates [20], which also capture epistemic uncertainty. On a pixel level, we compute standard uncertainty metrics: probability, entropy, predictive variance and mutual information (Appendix 5.2.4). On the image level, we want to obtain metrics to flag the prediction of a whole scene. As a simple heuristic, to account for spatial variability, we average the metrics over the (predicted) event size (Appendix 5.2.5). Discard-based evaluation allows to assess the practical utility of uncertainty estimates: we iteratively remove predictions with the highest estimated uncertainty and track segmentation performance (IoU, F1-score, accuracy) on the retained subset. Preliminary experiments show that the model is able to assign low confidence to predictions that are likely to be incorrect.

3 Results

Figure 2 shows our main outcomes: a) OOD metrics correlate with performance (samples are grouped by semantic HydroATLAS attributes), b) uncertainty flagging allows discarding lower-performing samples and c) visualizes both mechanisms integrated into a dashboard to help users make informed decisions on whether to trust predictions. A demo dashboard is online at `https://2025-esl-extreme-environments-demo-rj7dyok9w.vercel.app/` and further results are provided in Appendix 5.4. The results shown below are for MoCo (Appendix 5.3.1). To test the consistency of our findings, we evaluate SHRUG-FM across foundation models trained with different strategies (MAE, MoCo, DINO), all using a ViT-S/16 backbone (Appendix Table 2).

Out-of-Distribution Metric: As a first analysis, we grouped burn scar scenes by deciles of HydroATLAS features relevant to burn scar detection. For each group, we computed the mean NCDD and plotted it against the mean F1 score. We observe a robust linear relationship between NCDD and F1, implying that NCDD is an indicator of performance. Lower elevation or pasture extent and larger river area are associated with lower performance and higher OOD signals, indicating that predictions for these characteristics should be interpreted with caution. Fig. 5 in the Appendix visualizes these features and their performance on a map, highlighting a cluster of poorly performing scenes in the southeastern US. This aligns with low elevation, low pasture extent and high river area. While this is an initial, descriptive analysis, it suggests that foundation model developers may need to augment data this and similar regions. These initial outcomes underline the practical value of future systematic studies to identify features that consistently drive OOD behavior and to pinpoint undersampled areas. We also plan to examine cases where a scene is in-distribution in one space but OOD in the other (preliminary results are shown in Fig. 4 in the Appendix), which likewise warrant caution.

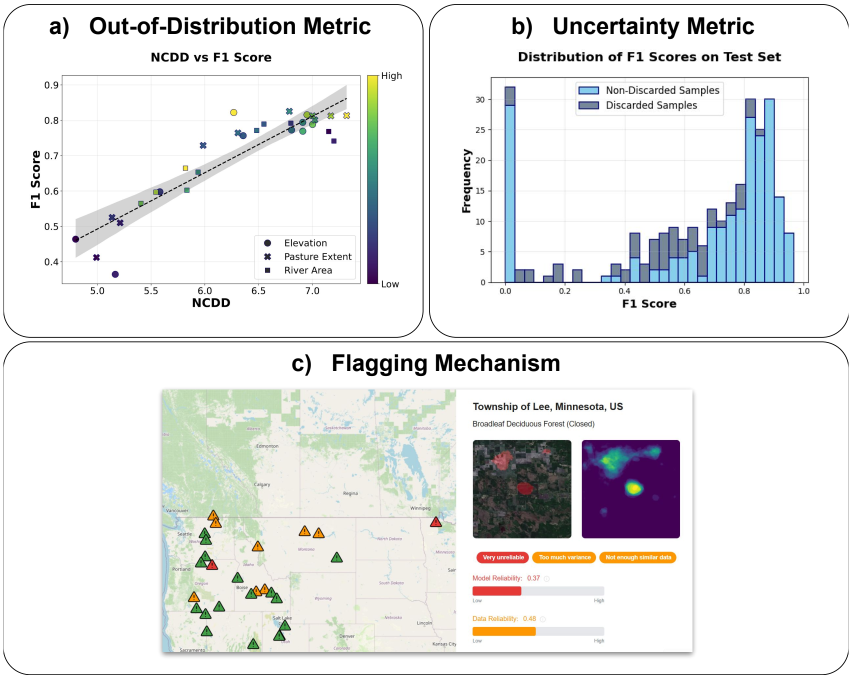

Figure 2: SHRUG-FM combines complementary signals to flag unreliable predictions. (a) The out-of-distribution (OOD) metric NCDD (in the embedding space) correlates with F1 scores (samples are grouped by HydroATLAS attributes). Low elevation, low pasture extent and large river areas are associated with lower performance and stronger OOD signals (higher NCDD) (b) A histogram of per-image test performance shows that the variance-based flag discards lower-performing samples. (c) Metrics are integrated into a dashboard, visualizing predictions, probability maps, reliability scores.

Uncertainty Metric: We applied an image-level discard flag based on the average predictive variance over the predicted burn scar, which is highest when (trained or dropout) ensemble members strongly disagree. By integrating over an adaptive region, this approach can also flag small scars with high uncertainty. The histogram in Fig. 2 shows that the flag successfully discarded many low-performing scenes while only affecting a few high-performing ones. However, it misses a set of poorly performing images where no burn scar was predicted. A less conservative threshold for the integration area, based on predicted probability, could help address this. Finally, the threshold that defines a failure mode should be determined by the downstream user, possibly requiring further calibration.

4 Conclusions and Pathway to Impact

SHRUG-FM offers practical steps to enhance the reliability of GFMs for EO in high-stakes environmental monitoring. By combining complementary uncertainty signals, related to the data (input and embedding OOD) and the downstream task (predictive uncertainty), it identifies where and why models may fail beyond standard benchmarks. The first important next step is to quantify the value of the proposed flags with metrics for discard mechanisms, like classification and rejection quality and the Area Under the Risk-Coverage Curve (AURC) [21]. We will also explore selective prediction mechanisms [22], training the system to combine the flags to abstain when uncertainty is high, going beyond setting a fixed threshold. To verify that SHRUG-FM works broadly, we will evaluate it across additional tasks, including flood mapping (with the WorldFloods dataset [23]) and landslide detection (with the CAS landslide dataset [24]) and across multiple GFMs (Prithvi [4], TerraMind [25]) that vary in pretraining strategy, model size and input modality. This comparative analysis will serve two purposes: first, to confirm that the flags consistently correlate with performance across tasks and models; and second, if this holds, to assess how different model designs and pretraining data choices influence reliability, guiding the building of more trustworthy GFMs in the future.

Acknowledgments and Disclosure of Funding

This work is a research product of ESL (ESL.ai), an initiative of the Frontier Development Lab, delivered by Trillium Technologies in partnership with ESA, University of Oxford, Google Cloud, and NVIDIA.

References

[1] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey, 2025. URL `https://arxiv.org/abs/2410.16602`.

[2] Clay Foundation Model. `https://madewithclay.org/`, 2024. Open-source geospatial foundation model website.

[3] 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 [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.

[4] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[5] Michael J Smith, Luke Fleming, and James E Geach. Earthpt: a time series foundation model for earth observation. *arXiv preprint arXiv:2309.07207*, 2023.

[6] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[7] Shan Zhao, Zhitong Xiong, Jie Zhao, and Xiao Xiang Zhu. Exebench: Benchmarking foundation models on extreme earth events, 2025. URL `https://arxiv.org/abs/2505.08529`.

[8] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. Pangaea: A global and inclusive benchmark for geospatial foundation models, 2025. URL `https://arxiv.org/abs/2412.04204`.

[9] 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 [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. doi: 10.1109/MGRS.2023.3281651.

[10] Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiao Xiang Zhu, and Tianjin Huang. Reobench: Benchmarking robustness of earth observation foundation models, 2025. URL `https://arxiv.org/abs/2505.16793`.

[11] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. Geo-bench: Toward foundation models for earth monitoring, 2023. URL `https://arxiv.org/abs/2306.03831`.

[12] Simon Linke, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, Vicente Burchard-Levine, Sally Maxwell, Hana Moidu, Florence Tan, and Michele Thieme. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data*, 6(1):283, Dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0300-6. URL `https://doi.org/10.1038/s41597-019-0300-6`.

[13] Juli G Pausas and Jon E Keeley. Wildfires and global change. *Frontiers in Ecology and the Environment*, 19(7):387–395, 2021.

[14] Peter F Moore. Global wildland fire management research needs. *Current Forestry Reports*, 5 (4):210–225, 2019.

[15] Alejandro Miranda, Rayén Mentler, Ítalo Moletto-Lobos, Gabriela Alfaro, Leonardo Aliaga, Dana Balbontín, Maximiliano Barraza, Susanne Baumbach, Patricio Calderón, Fernando Cárdenas, et al. The landscape fire scars database: mapping historical burned area and fire severity in chile. *Earth System Science Data*, 14(8):3599–3613, 2022.

[16] Sandesh Pokhrel, Sanjay Bhandari, Sharib Ali, Tryphon Lambrou, Anh Nguyen, Yash Raj Shrestha, Angus Watson, Danail Stoyanov, Prashnna Gyawali, and Binod Bhattarai. Ncdd: Nearest centroid distance deficit for out-of-distribution detection in gastrointestinal vision. *arXiv preprint arXiv:2412.01590*, 2024.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[18] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015. URL https://arxiv.org/abs/1412.1897.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[20] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

[21] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. *arXiv preprint arXiv:1805.08206*, 2018.

[22] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey, 2024. URL https://arxiv.org/abs/2107.11277.

[23] Enrique Portalés-Julià, Gonzalo Mateo-García, Cormac Purcell, and Luis Gómez-Chova. Global flood extent segmentation in optical satellite images. *Scientific Reports*, 13(1):20316, Nov 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-47595-7. URL https://doi.org/10.1038/s41598-023-47595-7.

[24] Yulin Xu, Chaojun Ouyang, Qingsong Xu, Dongpo Wang, Bo Zhao, and Yutao Luo. Cas landslide dataset: A large-scale and multisensor dataset for deep learning-based landslide detection. *Scientific Data*, 11(1):12, 2024.

[25] 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. *ICCV'25*, 2025.

[26] Qitian Ma, Shyam Nanda Rai, Carlo Masone, and Tatiana Tommasi. Segmentation re-thinking uncertainty estimation metrics for semantic segmentation, 2024. URL https://arxiv.org/abs/2403.19826.

[27] Tal Zeevi, Eléonore V. Lieffrig, Lawrence H. Staib, and John A. Onofrey. Spatially-aware evaluation of segmentation uncertainty, 2025. URL https://arxiv.org/abs/2506.16589.

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

[29] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

[31] Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Understanding contrastive versus reconstructive self-supervised learning of vision transformers. In *Self–Supervised Learning: Theory and Practice, Workshop at NeurIPS 2022*, December 2022.

[32] Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers, 2024. URL `https://arxiv.org/abs/2401.00463`.

[33] Gabriel Loaiza-Ganem, Valentin Villecroze, and Yixin Wang. Deep ensembles secretly perform empirical bayes, 2025. URL `https://arxiv.org/abs/2501.17917`.

[34] Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. *Workshop on Bayesian Deep Learning, NIPS*, 2016. URL `https://api.semanticscholar.org/CorpusID:8985844`.

5 Appendix

5.1 Data

5.1.1 SSL4EO

SSL4EO-S12 [9] is a large-scale dataset of multi-seasonal Sentinel-1 and Sentinel-2 (S2) imagery covering ~250,000 locations worldwide. The S2 data includes 13 spectral bands (B1–B8A, B9, B11, B12), preprocessed to a common 10m resolution. Locations are sampled throughout the year, ensuring seasonal diversity. This design captures temporal dynamics such as vegetation cycles, land cover change and climate-related variations. Together, these characteristics make SSL4EO-S12 one of the most comprehensive datasets available for large-scale self-supervised representation learning in Earth observation.

5.1.2 Burn scar segmentation: ExEBench

ExEBench Burn Scars [7] dataset consists of harmonized satellite imagery from Landsat and Sentinel-2 collected over burn scar areas between 2018 and 2021 over the United States. It contains 804 images of size 512×512, each with 6 spectral bands (covering visible, infrared, near-infrared and shortwave infrared). The resolution corresponds to 30 meters per-pixel.

5.1.3 Interpretable Attributes: HydroATLAS

HydroATLAS [12] is included in our experiments to provide complementary semantic analysis. The global BasinATLAS annotations of the HydroATLAS dataset provide hydro-environmental attributes spatially aggregated within a multi-level hierarchy of hydrological units. Our experiments use level 12 basins, the most granular resolution offered by HydroATLAS and the following characteristics, which are of particular relevance or of diagnostic value for the considered downstream tasks:

Table 1: BasinATLAS descriptions for the subset of attributes relevant for our downstream tasks.

Variable	**Description (Units)**
HYBAS_ID	HydroBasins ID
SUB_AREA	Sub-basin area extent (km^2)
inu_pc_smn	Inundation Extent (min annual), percent cover
inu_pc_smx	Inundation Extent (max annual), percent cover
ria_ha_ssu	River Area (hectares)
ele_mt_sav	Elevation (avg), meters a.s.l.
ele_mt_smn	Elevation (min), meters a.s.l.
ele_mt_smx	Elevation (max), meters a.s.l.
slp_dg_sav	Terrain Slope (avg), degrees (×10)
snw_pc_syr	Snow Cover Extent (avg annual), percent cover
snw_pc_smx	Snow Cover Extent (max annual), percent cover
glc_cl_smj	Land Cover Classes (spatial majority), classes (22)
wet_pc_sg1	Wetland Extent (class grouping 1), percent cover
wet_pc_sg2	Wetland Extent (class grouping 2), percent cover
for_pc_sse	Forest Cover Extent (avg), percent cover
crp_pc_sse	Cropland Extent (avg), percent cover
pst_pc_sse	Pasture Extent (avg), percent cover
ire_pc_sse	Irrigated Area Extent (avg), percent cover
urb_pc_sse	Urban Extent (avg), percent cover
hft_ix_s09	Human Footprint, index value (×10)

For each sample in the pretraining as well as the downstream task datasets, the aforementioned BasinATLAS attributes are queried at the patch's center in order to retrieve the values of the level 12 basin that the centroid falls in. This maps each data sample to a list of BasinATLAS attribute values.

5.2 Metrics

5.2.1 Out of Distribution Metrics

1. Nearest Centroid Distance Deficit (NCDD) For computing the NCDD values, we use the formulation described in [16].

$$NCDD = \alpha \cdot D_{others} - \beta \cdot D_{nearest}$$

where $D_{nearest}$ corresponds to the distance of a test point to the nearest centroid and D_{others} corresponds to the sum of its distances to all the other centroids. Since results were robust to different settings of α and β, we use α=1 and β=14 (where 14 corresponds to $k-1$). This further ensures that the metric is bounded between 0 and 14.

5.2.2 Performance Metrics

Metrics for Binary Predictions For a binary prediction (e.g., a pixel is either "burn scar" or "not burn scar"), we can evaluate performance using a confusion matrix with four key values: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).

1. IoU (Intersection over Union) Also known as the Jaccard Index, IoU measures the overlap between the predicted and ground truth areas. It is the ratio of the intersection of the two sets of pixels to their union.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

2. F1-score The F1-score is the harmonic mean of Precision ($\frac{\text{TP}}{\text{TP+FP}}$) and Recall ($\frac{\text{TP}}{\text{TP+FN}}$).

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

3. Accuracy Accuracy is the proportion of all correct predictions out of the total number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

For unbalanced tasks, where one class is significantly more frequent than the other (e.g., a small burn scar within a large image), Accuracy can be misleading. A model that simply predicts the majority class will achieve a high accuracy score, despite being useless for the minority class. In contrast, IoU and F1-score are more suitable as they focus on the positive class. Both metrics require the model to correctly identify a meaningful number of positive pixels to achieve a high score, making them robust to class imbalance.

5.2.3 Metrics for Predicted Probabilities

When models output probabilities, it is crucial that these probabilities are well-calibrated, meaning they accurately reflect the true likelihood of an event.

1. Expected Calibration Error (ECE) ECE measures how well a model's predicted probabilities align with the observed accuracy. The probability range is divided into a fixed number of bins and ECE is the weighted average of the absolute differences between the average predicted probability (confidence) and the actual accuracy within each bin.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

where M is the number of bins, $|B_m|$ is the number of samples in bin m, n is the total number of samples, $\text{acc}(B_m)$ is the accuracy in bin m and $\text{conf}(B_m)$ is the average confidence in bin m.

2. Adaptive Calibration Error (ACE) ACE is an improved version of ECE that addresses its limitations in handling unbalanced datasets. Instead of using equally-sized bins, ACE uses adaptively-sized bins so that each bin contains an approximately equal number of data points. This ensures that even for rare, low-confidence predictions, there are enough samples in a bin to provide a meaningful calibration estimate. For unbalanced datasets, many predictions may fall into a few high-confidence bins, which ECE may not accurately represent. By contrast, ACE's adaptive binning provides a more meaningful calibration score across the entire range of predictions, making it a more robust metric for unbalanced tasks.

5.2.4 Pixel-level Uncertainty Metrics

Given an ensemble of N models, each providing a probabilistic prediction $f_i(\mathbf{x}) = p_i$ for an input $\mathbf{x}$, we compute four key uncertainty measures.

1. Predicted Probability The predicted probability $\bar{p}$ is the average of the individual model predictions. It represents the ensemble's consensus on the most likely class and the estimated aleatoric uncertainty.

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$$

2. Predictive Entropy Predictive entropy $\mathcal{H}(\bar{p})$ measures the overall uncertainty of the ensemble's average prediction. A high value indicates a diffuse, uncertain prediction over multiple classes.

$$\mathcal{H}(\bar{p}) = -\sum_{c=1}^{C} \bar{p}_c \log(\bar{p}_c)$$

where C is the number of classes and $\bar{p}_c$ is the mean predicted probability for class c.

3. Predictive Variance Predictive variance $\mathbb{V}(p)$ quantifies the disagreement among the ensemble members. It captures the spread of individual model predictions.

$$\mathbb{V}(p) = \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2$$

4. Mutual Information Mutual information $\mathcal{I}$ measures the reduction in uncertainty about the predicted class provided by the ensemble. It captures the epistemic uncertainty or the uncertainty due to a lack of model consensus.

$$\mathcal{I} = \mathcal{H}(\bar{p}) - \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(p_i)$$

where $\mathcal{H}(p_i)$ is the entropy of the i-th model's prediction.

5.2.5 Image-level Uncertainty Metrics

1. Uncertainty over Region of Interest To obtain a single uncertainty value for an entire image, we aggregate a chosen pixel-level metric over a specific region of interest (ROI). For a given pixel-level uncertainty metric $\mathcal{M}(\mathbf{x})$ (e.g., Predictive Variance), the image-level uncertainty $\mathcal{U}_{image}$ is the average of that metric over the ROI.

$$\mathcal{U}_{image} = \frac{1}{|\text{ROI}|}\sum_{\mathbf{x}\in\text{ROI}} \mathcal{M}(\mathbf{x})$$

where $|\text{ROI}|$ is the number of pixels within the region of interest.

A common ROI is the predicted event area, defined as the set of pixels where the mean predicted probability $\bar{p}$ is greater than or equal to a threshold (e.g., 0.5 for a binary classification). Due to the clear interpretation and effectiveness, we chose the predictive variance over the predicted event area as the metric of choice for the results shown in Figure 2. In general, uncertainties highly depend on the task at hand and metrics specifically for borders and patches have been proposed [26, 27].

5.3 Methods

5.3.1 Foundation Model

The SSL4EO-S12 [9] dataset was published together with a suite of models trained on the data. This suite encompasses various established vision model architectures, including ResNets and ViTs of varying sizes, as well as trained weights from different pretraining strategies, such as MAE [28], DINO [29] and MoCo [30]. We chose for our analysis the ViT/s 16 model with ∼23 million parameters, which is a small vision transformer striking a balance between a performant state-of-the-art vision model and size for running comprehensive analyses. For further studies and our analysis, we focused on the MoCo ensemble as it was the model that yielded the best performance in F1 and IoU, as well as the best adaptive calibration error as shown in Table 2. Furthermore, it was found that contrastive methods such DINO or MoCo yield representation spaces that are better structured and separable making them more suitable for our approach to OOD detection [31, 32]. We hence proceed with the MoCo pretraining weights for all other baselines, such as the single model or the model with MC dropout.

5.3.2 Downstream Model

As a decoder, we chose a common deep CNN with ∼41 million parameters. By choosing an expressive decoder model, we ensure that if any problems in expressiveness arise, they would stem from the encoder foundation model as the focus should not be on the downstream model.

5.3.3 Downstream Uncertainty

We employ two established ensemble methods to estimate the aforementioned uncertainty metrics: Deep Ensembles [19] and Monte Carlo (MC) Dropout [20]. Both methods generate a set of diverse model predictions, from which we can compute predictive uncertainty.

1. Deep Ensembles Deep ensembles are a robust method for uncertainty estimation, as they have been shown to capture key properties of the Bayesian posterior distribution [33]. To create our ensemble, we train a set of 10 additional models with varying random seeds. To further enhance the diversity of the individual models and improve their uncertainty estimates in regions of low data density, each model is trained on a bootstrapped sample of the training data. [19] The main drawback of this approach is the high computational cost for both training and inference due to the need to train and store multiple full decoder models.

2. Monte Carlo Dropout As a more computationally efficient alternative, we utilize Monte Carlo (MC) Dropout [20]. This method approximates a Bayesian neural network by introducing dropout layers during training and keeping them active during inference. Dropout layers are inserted following non-linear activation functions within multi-layer convolution blocks, but not for the last component of a block. At test time, we perform 10 forward passes with dropout enabled to obtain a distribution of predictions from a single model. This approach is easy to implement and computationally cheaper than deep ensembles, as it does not require training multiple models. However, its expressiveness for capturing epistemic uncertainty can be limited, particularly for data within the training convex hull but in low-density regions [34].

5.4 Additional Results

Table 2: Performance of Baseline Models on Burn Scars Task

Model	F1 ↑	IoU ↑	Accuracy ↑	ECE ↓	ACE ↓
MoCo Ensemble	**0.8512**	**0.8532**	0.9021	0.0248	**0.0083**
MAE Ensemble	0.8499	0.852	**0.9041**	0.0274	0.0085
DINO Ensemble	0.831	0.8358	0.8934	0.0328	0.0109
MoCo Single	0.7278	0.7615	0.8282	**0.0102**	0.0348
MoCo Dropout	0.7957	0.8107	0.8805	0.0144	0.0181

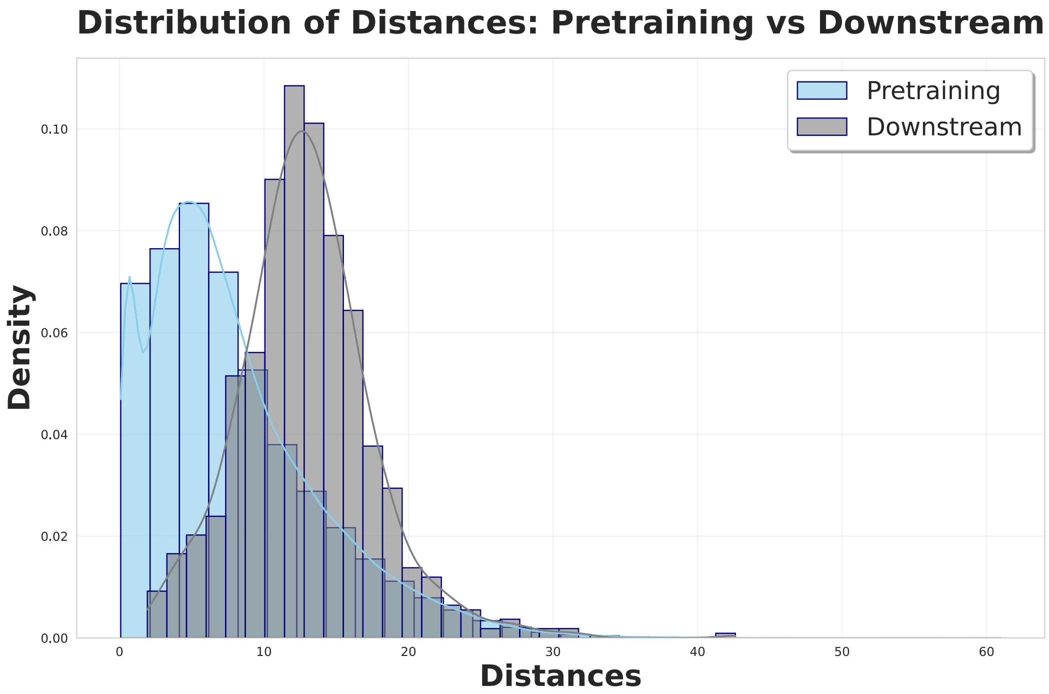

Figure 3: Density of distances to nearest k-means centroid for pretraining and downstream data. Downstream samples fall in higher-distance regions of the pretraining distribution (low density regions of the pretraining distribution).

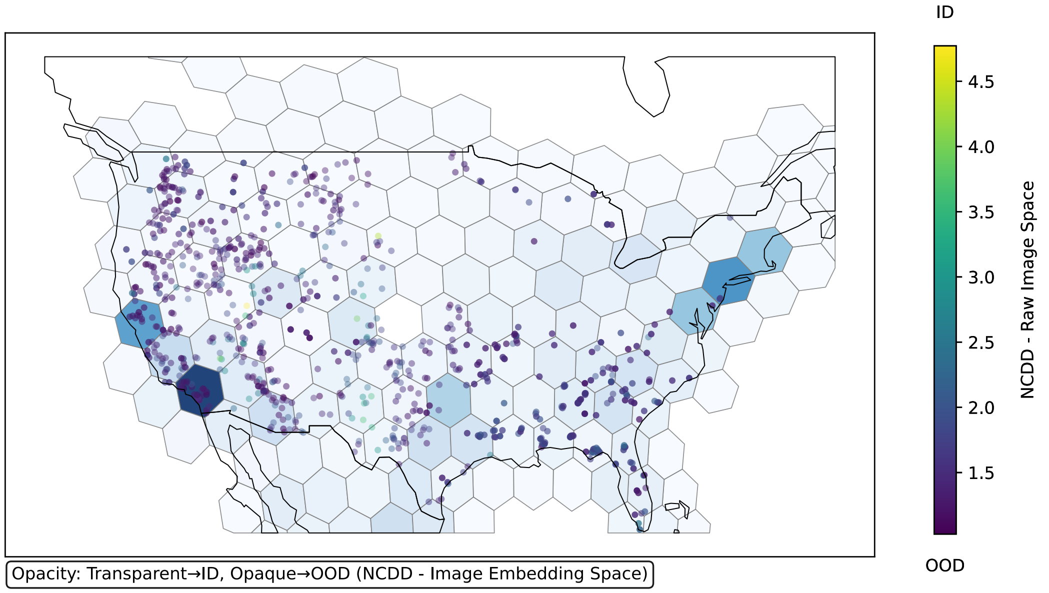
ID
4.5
4.0
3.5
3.0
2.5
2.0
1.5
NCDD - Raw Image Space
OOD
Opacity: Transparent→ID, Opaque→OOD (NCDD - Image Embedding Space)

Figure 4: Visualization of NCDD values for the downstream task overlaid on a hexagonal spatial distribution of SSL4EO-S12 across the US. Downstream data points are shown as scatter points positioned by their geographic locations. The color scale from purple to yellow represents increasing NCDD values in the raw image space, while the opacity of each point corresponds to NCDD values in the image embedding space. *Eg: An opaque point corresponds to a low NCDD value in the embedding space and a yellow colored point corresponds to a high NCDD value in the image space. This highlights potential concerns regarding sampling strategy employed during pretraining.*

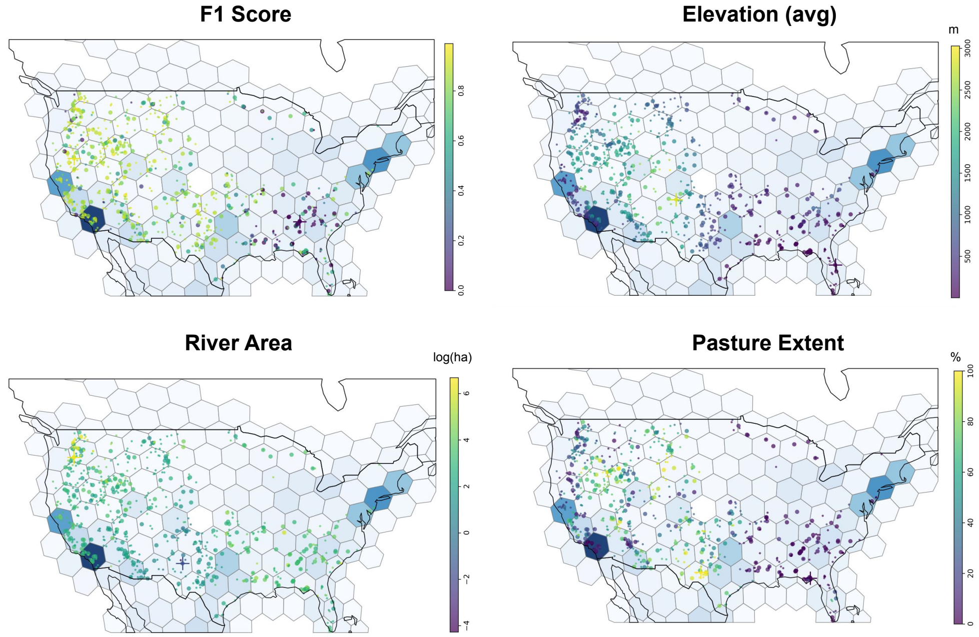

Figure 5: Visualization of F1 scores for the downstream task, Elevation (avg), River Area and Pasture Extent overlaid on a hexagonal spatial distribution of SSL4EO-S12 across the US. Cross (+) points represent highest/lowest values of the set. The F1 map highlights a low-performing cluster in the southeastern US, corresponding to a region characterized by low elevation, limited pastures and a relatively large river area. Moreover, the area contains few SSL4EO-S12 pretraining data points, hinting that the foundation model could benefit from increased data representation for this region.

Table 3: Performance Comparison of Different Scores

Score	**AUC Risk-Coverage ↓**	**Risk-Coverage at 0.5 ↓**	**AUC Nonrejected F1 ↑**	**Nonrejected F1 at 0.5 ↑**
Variance_0.2_scaled	0.345802	0.197721	0.630879	0.741266
Entropy_0.2_scaled	0.343770	0.186300	0.630627	0.743321
Mutual Information_0.2_scaled	0.345994	0.197721	0.631587	0.741184
normalized_distance_raw	0.525108	0.593164	0.605425	0.564644
-ncdd_raw	0.573665	0.570350	0.566803	0.549934
normalized_distance_embeddings	0.531331	0.494304	0.571290	0.603396
-ncdd_embeddings	0.386603	0.353603	0.704472	0.718206
-Elevation (avg)	0.322554	0.330789	0.723960	0.741955
-Pasture Extent	0.408957	0.315579	0.708769	0.755337
ria_ha_ssu	0.404475	0.357420	0.691717	0.718795
classifier_score	**0.211224**	**0.133067**	**0.756739**	**0.816070**
height				