

Adaptive Information Masking for Scalable Uncertainty Estimation in Geospatial Foundation Models

Mohammad Habibullah Rakib¹, Rupasree Dey¹, Tanjim Bin Faruk¹, Andrei Bachinin¹, F. Jay Breidt², Shrideep Pallickara¹, Sangmi Lee Pallickara¹

¹Colorado State University, Fort Collins, Colorado, USA

²NORC at the University of Chicago, Chicago, Illinois, USA

Abstract

Geospatial foundation models are applied to Earth-observation archives in which predictions inform high-stakes decisions and require reliable uncertainty estimates. Existing methods either train and store multiple models, or reuse a single trained model but recover better dropout masks only through a per-instance iterative search at inference. We present Adaptive Information Masking (AIM), a post-hoc extension of Monte Carlo (MC) Dropout that replaces iterative search with a single offline calibration. ZCA whitening decorrelates feature channels during calibration, and dropout masks are then modulated by saliency in the decorrelated space. On the Prithvi and DOFA geospatial foundation backbones, across landslide segmentation, land-cover classification and carbon-flux regression downstream tasks, AIM matches an iterative adaptive baseline on uncertainty quality without its inference-time optimization, and improves negative log-likelihood over uniform MC Dropout. Because no per-instance work remains at inference, AIM parallelizes across a multi-node cluster, which an iterative method cannot do without discarding the adaptivity that motivates it.

CCS Concepts

• Computing methodologies → Neural networks; • Information systems → Spatial-temporal systems; • Applied computing → Earth and atmospheric sciences.

Keywords

Uncertainty Quantification, Geospatial Foundation Models, Monte Carlo Dropout, Remote Sensing, Distributed Inference

ACM Reference Format:

Mohammad Habibullah Rakib, Rupasree Dey, Tanjim Bin Faruk, Andrei Bachinin, F. Jay Breidt, Shrideep Pallickara, and Sangmi Lee Pallickara. 2026. Adaptive Information Masking for Scalable Uncertainty Estimation in Geospatial Foundation Models. In *The 34th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26)*, November 03–06, 2026, Riverside, CA, USA. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3841645.3843066>

1 Introduction

Geospatial foundation models learn representations from large, heterogeneous collections of satellite imagery, climate records and

other spatiotemporal signals, and transfer them across tasks, regions and sensing modalities [7]. A prediction alone, however, is seldom sufficient. Earth systems are constantly changing, and distribution shifts frequently occur when models are deployed. Accurate estimation of epistemic uncertainty has become a critical aspect of model development, both for providing reliable information to support high-risk decision-making and for improving model performance [2, 13].

Two difficulties follow. Model uncertainty is entangled with the natural variability of the Earth system, since data distributions vary across geography, climate, season, sensor and land surface. Foundation models and the archives to which they are applied are also large, so a method that is accurate but computationally expensive remains impractical for Earth-observation workflows.

Existing approaches trade reliability against scalability. Deep ensembles provide strong estimates but require training and maintaining M models [10], and Bayesian neural networks are principled but difficult to deploy at foundation-model scale. MC Dropout requires only a single trained model, approximating Bayesian inference through repeated stochastic forward passes [3], but applies a uniform Bernoulli mask that drops every channel or location with the same probability. That assumption is hardly satisfied in learned representations, where spectral channels carry substantial redundancy. Adaptive dropout methods correct the uniformity by assigning per-layer rates from information-theoretic criteria [16], yet obtain those rates through an iterative search repeated for every input and every dropout rate, adding inference-time computation where deployment volume is highest. ZCA whitening addresses the same structure by removing cross-channel correlation [8].

We propose Adaptive Information Masking (AIM), a post-hoc extension of MC Dropout that requires no retraining and performs no per-instance optimization at inference, while keeping the masks conditioned on the input. This work makes three contributions. (1) AIM draws dropout masks in a ZCA-decorrelated feature space and modulates them by saliency, so that adaptivity follows from one offline calibration rather than from a per-instance search at inference. (2) We evaluate AIM, Rate-In [16] and uniform MC Dropout on Prithvi and DOFA across segmentation, classification and regression under a single matched protocol, holding the dropout rate and the dropout-site map identical, so that the differences are attributable to the masking mechanism alone. (3) We implement AIM as a distributed pipeline that pairs GPU calibration with CPU-parallel Spark inference, reaching a scalability factor of 8.3 at ViT-L scale, and we identify the point at which parallel efficiency saturates.



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGSPATIAL '26, Riverside, CA, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2950-8/2026/11

<https://doi.org/10.1145/3841645.3843066>

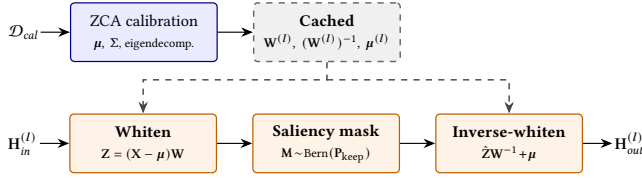


Figure 1: AIM at one dropout site.

Algorithm 1 Adaptive Information Masking (AIM)

Require: Trained network $\mathcal{F}$; dropout sites $\mathcal{S}$; calibration set $\mathcal{D}_{\text{cal}}$; base rate p ; MC iterations K ; test input $\mathbf{X}_{\text{test}}$; regularizer ϵ

Ensure: Stochastic predictions $\{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_K\}$

Phase 1: offline calibration ▶ run once, independent of p

- 1: **for** each site $l \in \mathcal{S}$ **do**
- 2: Collect activations $\mathbf{H}^{(l)}$ over $\mathcal{D}_{\text{cal}}$, reshape to $\mathbf{X} \in \mathbb{R}^{N \times C}$
- 3: $\boldsymbol{\mu}^{(l)} \leftarrow \text{mean}(\mathbf{X})$, $\tilde{\mathbf{X}} \leftarrow \mathbf{X} - \boldsymbol{\mu}^{(l)}$, $\boldsymbol{\Sigma}^{(l)} \leftarrow \frac{1}{N-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$
- 4: $\mathbf{U}\mathbf{A}\mathbf{U}^T \leftarrow \text{SVD}(\boldsymbol{\Sigma}^{(l)})$
- 5: Cache $\mathbf{W}^{(l)} \leftarrow \mathbf{U}(\mathbf{A} + \epsilon \mathbf{I})^{-1/2} \mathbf{U}^T$, $(\mathbf{W}^{(l)})^{-1}$, $\boldsymbol{\mu}^{(l)}$
- 6: **end for**

Phase 2: MC inference

- 7: **for** $k = 1$ to K **do**
- 8: **for** site $l \in \mathcal{S}$, input reshaped to $\mathbf{X}$ **do**
- 9: $\mathbf{Z} \leftarrow (\mathbf{X} - \boldsymbol{\mu}^{(l)}) \mathbf{W}^{(l)}$, $\tilde{\mathbf{S}} \leftarrow \text{minmax}_b(\text{mean}_c|\mathbf{Z}|)$
- 10: $\mathbf{P}_{\text{keep}} \leftarrow (1-p) + p\tilde{\mathbf{S}}$
- 11: $\mathbf{M} \sim \text{Bernoulli}(\mathbf{P}_{\text{keep}})$, $\hat{\mathbf{Z}} \leftarrow (\mathbf{Z} \odot \mathbf{M}) / \mathbf{P}_{\text{keep}}$
- 12: $\hat{\mathbf{X}} \leftarrow \hat{\mathbf{Z}} (\mathbf{W}^{(l)})^{-1} + \boldsymbol{\mu}^{(l)}$, reshape to native layout
- 13: **end for**
- 14: $\hat{\mathbf{y}}_k \leftarrow \mathcal{F}(\mathbf{X}_{\text{test}})$
- 15: **end for**
- 16: **return** $\{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_K\}$

2 Adaptive Information Masking

We formulate uncertainty estimation as a post-hoc procedure applied to an already trained model $\mathcal{F}_\theta$. AIM introduces stochasticity only at a fixed subset of layers $\mathcal{S}$ and assigns each site $l \in \mathcal{S}$ a base dropout rate p . Fixing $\mathcal{S}$ separates the masking mechanism from the choice of where dropout is placed. Site activations take one of three layouts, $\mathbb{R}^{B \times C \times h \times w}$ at convolutional sites, $\mathbb{R}^{B \times T \times C}$ at transformer sites and $\mathbb{R}^{B \times C}$ at fully connected sites, where B counts samples, C is the channel or embedding dimension, $h \times w$ the spatial resolution and T the number of tokens. Each layout exposes a channel dimension along which AIM estimates feature dependencies.

AIM proceeds in two phases, which Fig. 1 shows at a single site and Alg. 1 states in full. A calibration phase estimates a whitening matrix, its inverse and the channel mean at every site. This phase runs once, before deployment, and draws its data from $\mathcal{D}_{\text{cal}}$, a held-out subset of the training set. It updates no model weights and reads no test input. Each subsequent MC forward pass then whitens the activation, samples a saliency-modulated mask, and inverts the whitening.

Offline calibration. AIM reshapes the collected activations into $\mathbf{X} \in \mathbb{R}^{N \times C}$, treating each spatial location, token or sample as an independent observation of the C -dimensional channel vector, so

that $N = Bh w$, $N = BT$ or $N = B$ respectively. Off-diagonal entries of the covariance $\boldsymbol{\Sigma}^{(l)}$ identify correlated channels that tend to co-activate. Uniform dropout ignores these correlations. Dropping one channel of a correlated pair has little effect, since its partner carries redundant information, whereas dropping one of an uncorrelated pair removes information that nothing else supplies.

Zero-phase Component Analysis (ZCA) whitening removes these correlations. From the eigendecomposition $\boldsymbol{\Sigma}^{(l)} = \mathbf{U}\mathbf{A}\mathbf{U}^T$, AIM forms $\mathbf{W}^{(l)} = \mathbf{U}(\mathbf{A} + \epsilon \mathbf{I})^{-1/2} \mathbf{U}^T$, where ϵ is a regularization constant that ensures numerical stability. ZCA rotates the whitened data back to the original basis, unlike PCA, and so yields the representation closest to the original in the Euclidean sense [8], preserving the spatial alignment that the saliency step requires. AIM caches these three quantities per site as non-trainable parameters. Calibration costs $O(C^2 N_{\text{cal}} + C^3)$ operations per site and completes in seconds on a single GPU at $N_{\text{cal}} \approx 1920$. Because $\mathbf{W}^{(l)}$ depends entirely on feature covariance, one calibration serves every value of p .

Masking at inference. Whitening acts on the channel axis alone. AIM applies the cached $\mathbf{W}^{(l)}$ independently to every row of $\mathbf{X}$, so the transform never mixes information across locations, tokens or samples. Within the whitened space each channel carries unit variance and zero correlation with the others, so activation magnitude represents independent

AIM then computes the saliency map $\tilde{\mathbf{S}}$ by averaging absolute activation magnitudes across channels and normalizing the result to $[0, 1]$ within each sample, so that statistics never pool across the batch. The saliency modulates the keep probability as $\mathbf{P}_{\text{keep}} = (1-p) + p\tilde{\mathbf{S}}$, and the inverted scaling that AIM applies to the sampled mask preserves the expected activation, $\mathbb{E}[\hat{\mathbf{Z}}] = \mathbf{Z}$. A larger p therefore widens the gap between the keep probabilities of the least and most salient locations. AIM draws one mask value per location or token and broadcasts it across the C channels rather than drawing independently per element.

Inverse whitening returns the activations to the native layout and basis of the layer, so that subsequent layers receive inputs within the expected feature space. The projection also governs how the perturbation is distributed. Since $\tilde{\mathbf{S}} \geq 0$ implies $\mathbb{E}[\mathbf{P}_{\text{keep}}] \geq 1-p$, AIM zeroes fewer units than uniform dropout at a matched rate. The additional predictive variance arises instead from where the perturbation lands. Masking acts along decorrelated directions, and $(\mathbf{W}^{(l)})^{-1}$ spreads each suppressed direction across every native channel that loads on it, so a sparser mask in the whitened basis produces a broader, covariance-structured perturbation in the original basis.

3 Experimental Setup

Data. We evaluate on three downstream geospatial tasks. **Land-slide4Sense** [4] is binary landslide segmentation on 6-band HLS imagery at 128×128 (3,799/800 train/test), with landslide pixels under 2% of the total. **EuroSAT-MS** [5] is 10-class Sentinel-2 land-cover classification (21,600/2,700), resized $64 \rightarrow 224$ for ViT compatibility, with the 13 Sentinel-2 bands mapped onto the 6-band HLS convention shared by both backbones. **CarbonFlux** [6] regresses standardized Gross Primary Productivity from HLS-6 chips at 224×224 (688/287). To test whether the mechanism generalizes beyond Earth observation, we additionally evaluate on CIFAR-100

[9], three MedMNIST v2 (BloodMNIST, PathMNIST, TissueMNIST) medical imaging datasets [15] at both 28×28 and 224×224, and three UCI tabular regression (Protein, Wine, Energy) benchmarks [1].

Backbones and dropout placement. Prithvi-EO-2.0-300M [12] and DOFA-Large [14] are 24-block ViT-L masked-autoencoder backbones with $C = 1024$, pretrained respectively on 6-band HLS imagery and on multi-sensor imagery with dynamic-wavelength embedding. Every compared method uses the identical site map $\mathcal{S}$, namely blocks $\{0, 6, 12, 18, 23\}$ with both `attn.proj_drop` and `m1p.drop2` per block ($|\mathcal{S}| = 10$) and all other dropout zeroed, so that only the masking mechanism varies. On the non-geospatial backbones of Sec. 4, placement follows each architecture: we inject AIM after the convolutional blocks of ResNet-20/56, RepVGG-A1/A2 and ResNet-18 (4D activations), and replace the existing fully connected dropouts of VGG-16/19-BN and of the regression multilayer perceptrons (2D activations).

Protocol. All foundation-model results use a matched base rate $p = 0.1$, $K = 20$ Monte Carlo samples and three seeds, reported as mean $\pm$ std, so that no method receives a test-set-selected operating point. The exception is the non-geospatial regression benchmarks, for which predictive variance is strongly p -sensitive and each method is reported at its lowest-NLL rate. Only the modules in $\mathcal{S}$ switch to training mode, so that normalization statistics stay frozen and identical across methods, and calibration draws on training data only. GPU runs use one NVIDIA A100 3g.40gb MIG partition under PyTorch 1.12, and distributed inference uses a 23-node CPU cluster running Apache Spark.

Metrics. For classification we report Expected Calibration Error (ECE, 15 bins) and Negative Log-Likelihood (NLL) for calibration, the Area Under the Accuracy-Rejection Curve (AUARC) for selective prediction, and top-1 Accuracy (ACC) for task quality, substituting macro-Jaccard on segmentation, where class imbalance makes pixel accuracy uninformative. For regression we report Gaussian NLL, Predictive Interval Coverage Probability (PICP) at 95% nominal coverage, the Interval Efficiency Ratio (IER) for interval sharpness relative to coverage, and Mean Squared Error (MSE). The aggregate uncertainty score is the mean predictive entropy of the MC ensemble, and its epistemic component is the mutual information (MI) between predictions and parameters.

4 Results

Uncertainty quality. We compare AIM against Rate-In [16], a representative iterative adaptive method, and against the uniform-mask baseline (Table 1). Three of the four settings are saturated, by class imbalance on Landslide4Sense and by the Sentinel-2 pre-training of DOFA on EuroSAT, so accuracy and AUARC carry little information there and the methods separate only on calibration. Only EuroSAT under Prithvi is unsaturated, where HLS pretraining leaves accuracy between 57% and 62% and no method dominates.

A single pattern runs through all four settings. AIM is the stronger absolute-probability calibrator, best or tied on NLL everywhere and best on ECE in two settings, whereas Rate-In holds the best selective-risk ranking in the one unsaturated setting, by less than one standard deviation (σ). Task quality separates no method. The segmentation measures we report outside Table 1, macro-Jaccard

Table 1: Uncertainty quality on geospatial foundation models at matched $p = 0.1$, $K = 20$, three seeds. Best per group in bold. For Landslide4Sense, ACC denotes pixel accuracy.

Method	ACC $\uparrow$	NLL $\downarrow$	ECE $\downarrow$	AUARC $\uparrow$
<i>Landslide4Sense Segmentation, Prithvi</i>				
Baseline	0.982	0.116 $\pm$.004	0.076 $\pm$.003	0.999
AIM	0.983	0.109 $\pm$.005	0.071 $\pm$.005	0.999
Rate-In	0.982	0.124 $\pm$.005	0.070 $\pm$.005	0.999
<i>Landslide4Sense Segmentation, DOFA</i>				
Baseline	0.983	0.118 $\pm$.010	0.078 $\pm$.010	0.999
AIM	0.982	0.118 $\pm$.009	0.075 $\pm$.008	0.999
Rate-In	0.982	0.119 $\pm$.010	0.078 $\pm$.009	0.999
<i>EuroSAT-MS Classification, Prithvi</i>				
Baseline	0.566 $\pm$.051	1.305 $\pm$.155	0.148 $\pm$.035	0.770 $\pm$.054
AIM	0.617 $\pm$.006	1.090 $\pm$.049	0.108 $\pm$.035	0.832 $\pm$.004
Rate-In	0.615 $\pm$.008	1.108 $\pm$.091	0.128 $\pm$.052	0.835 $\pm$.006
<i>EuroSAT-MS Classification, DOFA</i>				
Baseline	0.961 $\pm$.001	0.357 $\pm$.018	0.030 $\pm$.001	0.988 $\pm$.001
AIM	0.966 $\pm$.001	0.345 $\pm$.026	0.031 $\pm$.001	0.989 $\pm$.000
Rate-In	0.964 $\pm$.001	0.370 $\pm$.033	0.032 $\pm$.002	0.988 $\pm$.002

Table 2: CarbonFlux regression. NLL_{cal} follows one global variance recalibration fitted on validation.

Backbone	Method	MSE $\downarrow$	NLL _{cal} $\downarrow$	PICP $\uparrow$	IER $\downarrow$
Prithvi	Baseline	0.160	0.579	0.182	0.884
	AIM	0.151	0.812	0.382	0.700
	Rate-In	0.179	0.341	0.196	0.770
DOFA	Baseline	0.150	0.337	0.074	0.691
	AIM	0.145	0.495	0.220	0.673
	Rate-In	0.148	0.326	0.070	0.724

at ≈ 0.61 and landslide-class IoU between 0.23 and 0.25, fall within 1σ as well.

Regression. On CarbonFlux (Table 2) AIM obtains the best MSE on both backbones, with the remaining differences within noise. Raw MC Dropout is severely under-dispersed on this target, as 95%-nominal coverage collapses to between 0.07 and 0.38 for every method, because the ensemble captures epistemic but not aleatoric variance. AIM comes closest to nominal, at two to three times the baseline coverage. This is the behavior that the inverse whitening of Sec. 2 implies, since a sparser mask in the decorrelated basis produces a broader perturbation in the original one. Since the raw predictive variance is uncalibrated for all three methods, we report NLL after a single global rescaling of the predictive variance, fitted on the validation split and applied identically to each, and Rate-In is best on that measure. The split mirrors the classification result. AIM widens and sharpens the predictive interval, while Rate-In produces the better-scaled likelihood.

Computational cost. AIM separates calibration from inference. Calibration requires under a minute at ViT-L scale and is reused

across dropout rates, whereas Rate-In repeats a per-layer MI optimization for every rate, so the overhead of AIM is constant rather than linear in the number of operating points evaluated. Its only inference-time addition is two $C \times C$ products per active site.

Distributed inference. Parallel MC inference is available to any forward-only method, so the speedup alone does not distinguish AIM. However, AIM is both forward-only and per-instance adaptive, so distribution preserves its adaptivity, whereas Rate-In obtains its rates through an iterative forward-and-backward optimization that runs one to two orders of magnitude slower than a forward pass on CPU, and distributes identically only by freezing those rates and thereby discarding the adaptivity that motivates it. Because the cached statistics are static and each mask depends only on its own input, inference parallelizes over test instances without inter-worker communication. Our Apache Spark implementation calibrates once on a GPU node, writes the checkpoint to a shared file system storage and broadcasts only partition indices, so that each worker loads the checkpoint and its own shard directly and neither model weights nor imagery pass through the driver. Writing $T(N)$ for the MC-inference wall-clock over the full test set on N worker instances, scaling is characterized by the scalability factor $S(N) = T(1)/T(N)$ (ideally N) and the parallel efficiency $E(N) = S(N)/N$. Over $N \in \{1, 2, 4, 8, 12, 16, 20, 46\}$ instances on 23 nodes, Prithvi attains a maximum $S = 8.3$ at $N \in [16, 20]$. Beyond $N = 20$, E degrades from ≈ 0.5 to ≈ 0.14 , as the per-partition workload falls below the subprocess and thread-contention overhead under full two-instance co-location.

Beyond geospatial backbones. Across CIFAR-100 with six CNN backbones, three MedMNIST datasets at two resolutions, and three UCI regression benchmarks, the advantage of AIM concentrates where dropout is placed on high-dimensional fully connected layers. On the two VGG variants it approximately halves ECE (0.194 to 0.058 and 0.185 to 0.079), whereas on small spatial-injection backbones it reduces to parity and the baseline retains the lowest ECE and NLL. On regression AIM obtains the best NLL and IER on five of six dataset-architecture combinations. Decorrelation yields benefit in proportion to the covariance structure available at the dropout site, which is greatest in the fully connected regime. These smaller backbones also distribute further, reaching $S \approx 20$ where the ViT-L models saturate at 8.3.

5 Conclusion and Limitations

AIM estimates uncertainty in geospatial foundation models without retraining, ensembling or per-instance optimization. A single offline calibration drives dropout maps weighted by information content. Across Prithvi and DOFA it reduces NLL by roughly 6% on average over the uniform baseline and by 16% in the least saturated setting, while task quality and selective-risk ranking remain unchanged. Because the calibration overhead is incurred once offline, inference distributes without repeating the adaptive step, reaching a scalability factor of 8.3 at foundation-model scale.

The gains concentrate in absolute-probability calibration, and where the task saturates all methods sit within noise. On regression the predictive intervals of every method, AIM included, remain far

below nominal coverage. AIM is a structured perturbation estimator rather than a posterior approximation. Its mask is conditioned on the input, so the variational argument that justifies MC Dropout as approximate Bayesian inference [3] does not carry over, and we validate it empirically through calibration and selective risk. Deep ensembles and approximate-posterior methods such as SWAG [11] define the quality ceiling for predictive uncertainty. AIM addresses the complementary regime in which their training and storage requirements rule them out. Separating the contribution of whitening from that of saliency modulation, through masking in the original basis, is the natural next step.

Acknowledgments

This research was supported by the National Science Foundation (2312319, 2617041), the National Institute of Food Agriculture (COL0-14021223, 2025-77039-45531), an NSF/NIFA Artificial Intelligence Institutes AI-LEAF Award [2023-03616], and a Clare Boothe Luce Professorship.

References

- [1] Stephen D Bay, Dennis Kibler, Michael J Pazzani, and Padhraic Smyth. 2000. The UCI KDD archive of large data sets for data mining research and experimentation. *ACM SIGKDD explorations newsletter* 2, 2 (2000), 81–85.
- [2] Rupasree Dey, Abdul Matin, Nathan Orwick, Yao Zhang, Shrideep Pallickara, and Sangmi Lee Pallickara. 2026. When to Trust, How to Distill: Multi-foundation model guidance for lightweight, robust scientific time series forecasting. In *Proceedings of the 32nd ACM SIGKDD*. 10807–10818.
- [3] Yarín Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059.
- [4] Omid Ghorbanzadeh, Yonghao Xu, Pedram Ghamisi, et al. 2022. Landslide4Sense: Reference Benchmark Data and Deep Learning Models for Landslide Detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–17. doi:10.1109/TGRS.2022.3215209
- [5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226. doi:10.1109/JSTARS.2019.2918242
- [6] IBM Geospatial and NASA IMPACT. 2024. HLS MERRA2 GPPFlux Dataset. https://huggingface.co/datasets/ibm-nasa-geospatial/hls_merra2_gppflux. Accessed: June 2026.
- [7] Johannes Jakubik, Sujit Roy, CE Phillips, et al. 2023. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660* (2023).
- [8] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. Optimal Whitening and Decorrelation. *The American Statistician* 72, 4 (2018), 309–314.
- [9] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).
- [11] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems* 32 (2019).
- [12] Daniela Szwarzman, Sujit Roy, Paolo Fraccaro, et al. 2026. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. *arXiv:2412.02732 [cs.CV]* <https://arxiv.org/abs/2412.02732>
- [13] Helin Wang, Wei Du, Ning Liu, Qian Li, Yanyu Xu, and Lizhen Cui. 2025. AdaHetMKD: An Adaptive Heterogeneous Multi-teacher Knowledge Distillation for Medical Image Analysis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.
- [14] Zhitong Xiong, Yi Wang, Fahong Zhang, et al. 2025. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. *arXiv:2403.15356 [cs.CV]*
- [15] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* 10, 1 (2023).
- [16] Tal Zeevi, Ravid Shwartz-Ziv, Yann LeCun, et al. 2025. Rate-In: Information-Driven Adaptive Dropout Rates for Improved Inference-Time Uncertainty Estimation. In *2025 IEEE/CVF CVPR*. 20757–20766. doi:10.1109/CVPR52734.2025.01933