Self-Supervised Multi-Modal World Model with 4D Space-Time Embedding

Lance Legel*
Ecological Intelligence Lab
Ecodash.ai

Qin Huang
School of Complex Adaptive Systems
Arizona State University

Brandon Voelker
Geosensing Systems Engineering & Sciences Lab
University of Houston

Daniel Neamati
Navigation & Autonomous Vehicles Lab
Stanford University

Patrick Alan Johnson
Earth System Lab
Allen Institute for Artificial Intelligence

Favyen Bastani
Earth System Lab
Allen Institute for Artificial Intelligence

Jeff Rose
Spatial Intelligence Lab
SpatialLogic.com

James Ryan Hennessy
Department of Computer Science
Georgia Tech Institute of Technology

Robert Guralnick
Florida Museum of Natural History
University of Florida

Douglas Soltis
Florida Museum of Natural History
University of Florida

Pamela Soltis
Florida Museum of Natural History
University of Florida

Shaowen Wang
NSF Institute for Geospatial Understanding
University of Illinois Urbana-Champaign

Abstract

We present *DeepEarth*, a self-supervised multi-modal world model with *Earth4D*, a novel planetary-scale 4D space-time positional encoder. Earth4D extends 3D multi-resolution hash encoding to include time, with the potential to efficiently scale across the planet over centuries with sub-meter, sub-second precision. Multi-modal encoders (*e.g.* vision-language models) are fused with Earth4D embeddings and trained via masked reconstruction. We demonstrate Earth4D's expressive power by achieving state-of-the-art performance on an ecological forecasting benchmark. Earth4D with learnable hash probing surpasses a multi-modal foundation model pre-trained on substantially more data. Access open source code and download models at: `https://github.com/legel/deepearth`.

1 Introduction

Revise introduction for world modeling audience (vs ecology audience) (Reviewer [VajQ])

Clarify distinction between world model and neural field in introduction (Reviewer [QJfa])

Expand literature review in introduction (Reviewers [QS1P], [QJfa])

*Correspondence: `lance@ecodash.ai`

Discuss dependence on coordinates and species labels alone and its limitations (Reviewer [QS1P])

More clarity on multimodality (Reviewer [QJfa])

As learning-based techniques expand into the real world, a broad area of disciplines feature related challenges of bringing geospatial data at sparse sites collected at irregular intervals into a unified framework. World models frame the learning problem as an interactive opportunity to consider counterfactuals or forecast future events, made even more pressing with the growing of artificial intelligence agents.

Cite: `https://www.nvidia.com/en-us/glossary/world-models/`

In the context of land management (including, but not limited to, ecology, landscaping, or agriculture), decision-makers have the opportunity to weigh options based on the world model's predictions. For example, predicting the timing of flowering or live fuel moisture content across scales can inform downstream decisions, such as when and where to plant specific species or take mitigating actions against natural hazards.

Many world models are built for artificial intelligence agents, though not necessarily for land management. review on types of world models

reference multi-modal world models

Recent world models have considered geospatial data. context of Galileo and AlphaEarth, since used later in the paper

We contrast this development of world models with the development of neural fields. These representations consider a maping of coordinate-based features to a physical value. For example, a neural gravitational field would map spatial coordinates to a gravitational potential . Similar representations can be constructed for other physical fields, such as for glacial flows , . Many of these neural fields use physics-informed neural networks (PINNs) to constrain the learned mapping where known physical laws are directly included in the learning objective. However, in many settings in land management, such as ecological forecasting, clean physical laws are not easily available. One approach to learn neural fields is to use self-supervised learning, where the model is trained to reconstruct the input data. This approach has been successfully applied to weather forecasting , . Implicitly, the domain knowledge (e.g., meteorology) is built into the input-output pairs or model architecture and the method succeeds due to the availability of large amounts of data.

cite

cite

finish

cite

finish

A particular type of neural field that motivates this work is the neural radiance field (NeRF), which maps spatial coordinates and view directions to a color or density contribution . The physics-based volumetric rendering constrains how the NeRF must allocate color or density contributions to the scene, to reconstruct the input training images. Early works in NeRF were extremely time-consuming to train, requiring hours or days on a single GPU . Although many works aimed to reduce the training time , the introduction of instant neural graphics primitives (INGPs) (Müller et al., 2022) and multi-resolution hash encoding (MHE) (Takikawa et al., 2023) allowed for the training of NeRFs on small batches in minutes . 3D reconstruction is typically considered for small scenes (e.g., on the scale of homes), but recent works have demonstrated the ability to reconstruct large scenes (e.g., on the scale of cities) . Other works have also considered aerial or satellite imagery and have explored the ability to reconstruct across devices (e.g., ground and aerial) in a georeferenced manner . Although some works have considered sensors beyond camera imagery (e.g.,), These 3D reconstructions are effective and useful in the context of land management, but they constrain the type of data that can be stored and reconstructed. Fundamentally, the core of 3D reconstruction and neural fields in computer vision, is to recover the training images and generalize to new camera viewpoints or lighting conditions, rather than handle sparse geospatial data.

cite

cite

cite

verify citations

cite

cite

cite

cite

Our key insight is to reconsider what data could be effectively stored in a multi-resolution hash encoding and neural field that would elevate the performance of world models in land management, al-

lowing more compact models and improved performance. We present DeepEarth, a self-supervised multi-modal world model that takes an interdisciplinary approach to efficiently and effectively represent geospatial data. Although applicable to a wide range of domains, we evaluate DeepEarth
on three tasks: live fuel moisture content prediction, RGB aerial imagery reconstruction, and hy- check
drological forecasting. Live fuel moisture content (LFMC) prediction is a key indicator of timely wildfire risk, impacting mitigation strategies before a fire starts as well as supporting fire suppression opertaions (e.g., where to deploy resources and which fuels to target). For LFMC prediction, we evaluate DeepEarth on the Globe-LFMC 2.0 ecological forecasting benchmark that separates LFMC by species allowing for land managers and AI agents to consider the how a species planting choice can impact future fire risk. For RGB aerial imagery reconstruction, we evaluate DeepEarth on the
USGS 3DEP LiDAR and USDA NAIP imagery paired by Allred et al. to reconstruct RGB pixels cite
from a timestamped geolocated point cloud. This reconstruction task, similar to NeRF, addresses to what extent the visual appearance of the scene is accurately mapped to geolocated point queries.
For hydrological forecasting, we evaluate DeepEarth on . These tasks illustrate the verisitility of finish
DeepEarth across domains and the potential of world models in land management.

2 DeepEarth Architecture

DeepEarth is a self-supervised multi-modal world model that learns unified representations of Earth observation data across space and time. As seen in Figure 1, the architecture processes multi-modal inputs (*e.g.* vision, language, sensor data) sampled around spatio-temporal events. The Earth4D encoder maps continuous space-time coordinates (*latitude*, *longitude*, *elevation*, *time*) to learnable positional embeddings, which are fused with embeddings from modality-specific encoders and processed as tokens in an autoencoder context window. Inspired by PerceiverIO (Jaegle et al., 2021), V-JEPA 2 (Assran et al., 2025), Galileo (Tseng et al., 2025), and AlphaEarth (Brown et al., 2025), DeepEarth learns to generatively reconstruct and simulate joint distributions of multi-modal data.

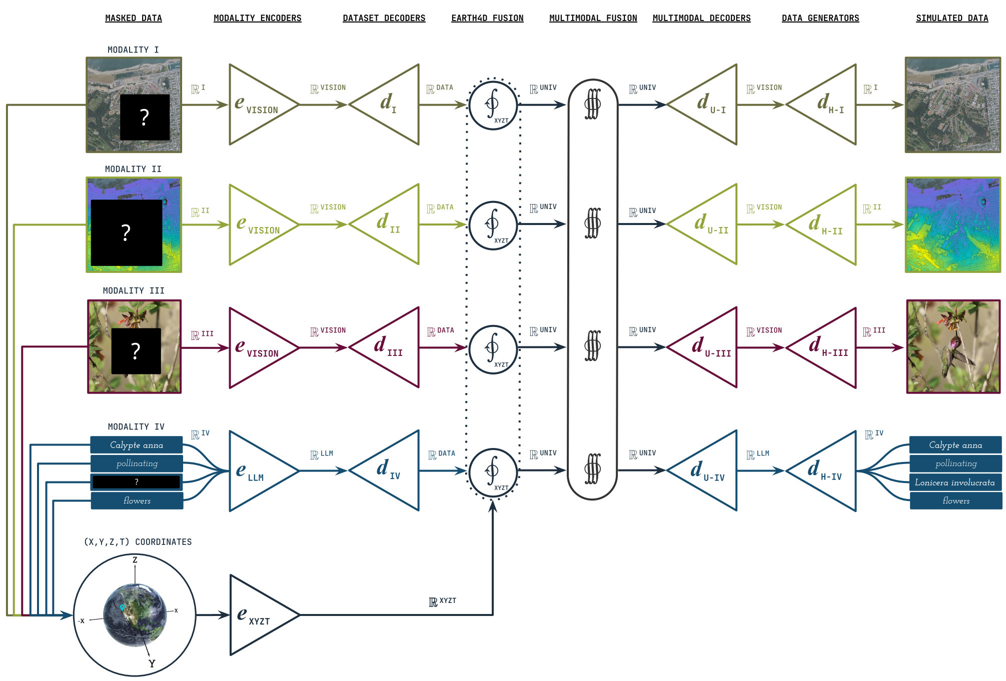

Figure 1: **DeepEarth Overview.** Masked multi-modal data (*e.g.* images, text) sampled around an event (*e.g.* pollination) are encoded and fused with Earth4D space-time embeddings. These universal tokens are jointly encoded, and then masked data is inductively decoded and simulated.

3 EARTH4D ARCHITECTURE

Following Grid4D (Jiawei et al., 2024), Earth4D extends NVIDIA's multi-resolution hash encoding (Müller et al., 2022) to four dimensions (Figure 2) by concatenating features from one spatial (*xyz*) and three spatio-temporal (*xyt*, *yzt*, *xzt*) grids. Implemented as a standalone PyTorch module with massively parallelizable CUDA kernels, Earth4D is suitable for integration into larger models.

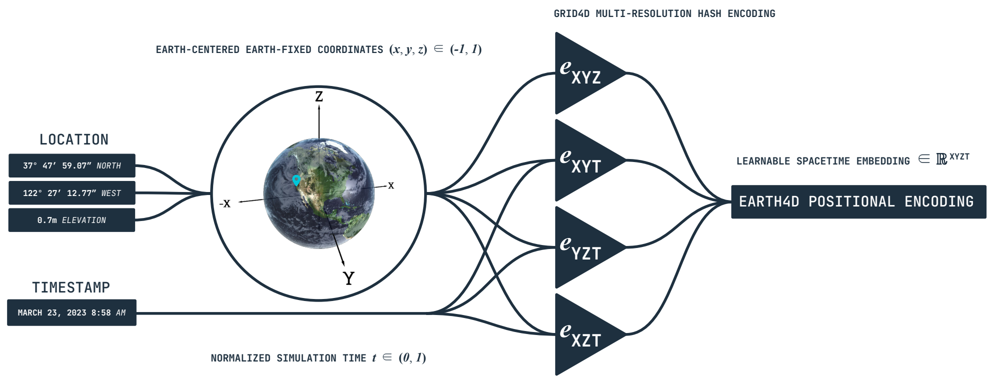

Figure 2: **Earth4D Space-Time Positional Encoding.** A planetary-scale 4D encoder with fully decomposable spatio-temporal representation. Four grids (*xyz*, *xyt*, *yzt*, *xzt*) are each learned in 3D space and computed in parallel. Each grid has multiple resolution levels (Appendix A), enabling deep learning of complex joint distributions in multi-modal data across space-time scales.

Earth4D's hash encoding compresses spatial features into a fixed memory budget, but different coordinates can map to the same memory location (collisions). We integrate learned hash probing (Takikawa et al., 2023), an end-to-end differentiable system that learns optimal memory allocation patterns for the data. This yields substantial performance improvements across tasks (Appendix B).

4 EARTH4D EXPERIMENTAL VALIDATION

4.1 LIVE FUEL MOISTURE CONTENT PREDICTION

Dataset. Live Fuel Moisture Content (LFMC) measures the percentage of water in vegetation relative to its dry weight, a critical indicator for wildfire risk assessment. We evaluate Earth4D on Globe-LFMC 2.0 (Yebra et al., 2024), a global ecological forecasting benchmark containing field measurements across diverse plant species, geographic regions, and temporal periods.

Baseline Model. We compare against Galileo (Johnson et al., 2025; Tseng et al., 2025), a pre-trained Vision Transformer processing multi-modal remote sensing data (Appendix C).

Architecture. Earth4D encodes (x,y,z,t) into a 192D vector, concatenated with a learnable species embedding initialized randomly (no prior knowledge). An MLP then predicts LFMC %.

Results. Earth4D achieves MAE 12.1pp and R^2 0.755, surpassing Galileo (MAE 12.6pp, R^2 0.72) using only (x,y,z,t) coordinates and species embeddings (Table 1).

Model	Data Inputs	MAE (pp)	RMSE (pp)	R^2
Galileo (Pre-Trained)	(x,y,z,t) + Species Type + Remote Sensing	12.6	**18.9**	0.72
Earth4D (Learned Hash)	(x,y,z,t) + Species Name	**12.1**	19.9	**0.755**

Table 1: **State-of-the-Art Ecological Forecasting Benchmark.** Earth4D surpasses the pre-trained Galileo foundation model without satellite imagery, weather data, or topography.

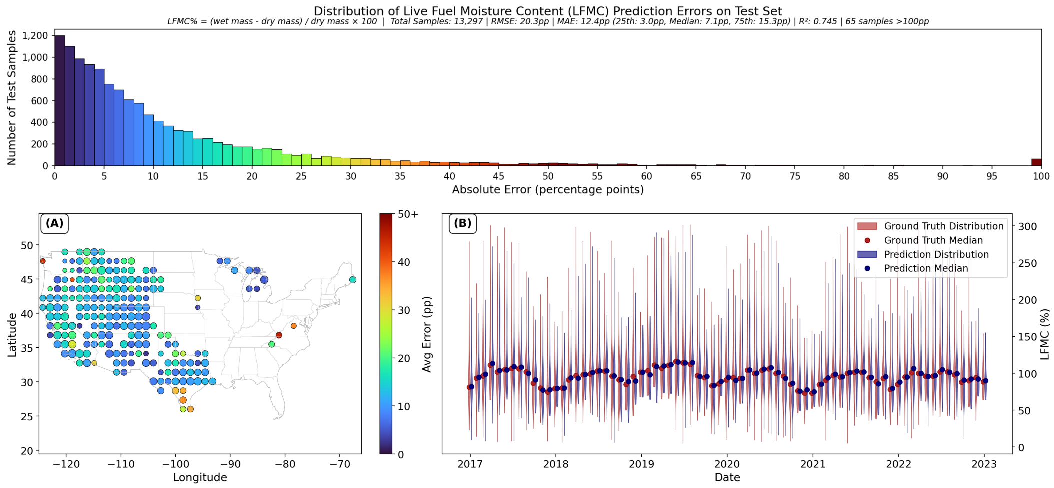

Figure 3: **Earth4D LFMC Prediction Performance.** *(Top)* Distribution of absolute errors in percentage point predictions across 13,297 test samples, showing median error of 7.1pp. *(A)* Geographic error distribution across CONUS shows low error in well-sampled regions. *(B)* Temporal predictions closely track ground truth LFMC measurements across seasons (2017–2023).

4.2 RGB Aerial Imagery Reconstruction

We evaluate Earth4D's ability to infer RGB pixels from (x,y,z,t) inputs with objective $(x,y,z,t) \rightarrow (r,g,b)$. Using USGS 3DEP LiDAR (Stoker & Miller, 2022; Sugarbaker et al., 2014) and USDA NAIP imagery (USDA, 2003–present) paired by Allred et al. (2025), we train on 5.8M coordinate-color pairs from Houston coastal wetlands (Figure 4).

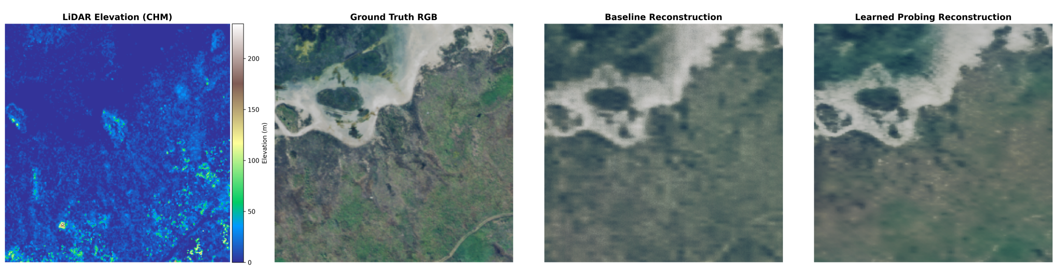

Figure 4: **RGB Reconstruction from LiDAR Elevation.** Houston coastal wetlands, 2018. *Left to right:* LiDAR height, ground truth, baseline, learned probing (18% lower loss).

References

John T Abatzoglou, Solomon Z Dobrowski, Sean A Parks, and Katherine C Hegewisch. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data*, 5:1–12, 2018.

Brady W Allred, Sarah E McCord, and Scott L Morford. Canopy height model and NAIP imagery pairs across CONUS. *Scientific Data*, 12(1):322, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

Tom G Farr and Mike Kobrick. Shuttle radar topography mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81:583–585, 2000.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2021.

Xu Jiawei, Fan Zexin, Yang Jian, and Xie Jin. Grid4D: 4D decomposed hash encoding for high-fidelity dynamic gaussian splatting. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Patrick Alan Johnson, Gabriel Tseng, Yawen Zhang, Heather Heward, Virginia Sjahli, Favyen Bastani, Joseph Redmon, and Patrick Beukema. High-resolution live fuel moisture content (LFMC) maps for wildfire risk from multimodal earth observation data, 2025. URL https://arxiv.org/abs/2506.20132.

Joaquín Muñoz Sabater. ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2019.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, July 2022. ISSN 1557-7368. doi: 10.1145/3528223.3530127. URL http://dx.doi.org/10.1145/3528223.3530127.

Jason Stoker and Barry Miller. The accuracy and consistency of 3D elevation program data: A systematic analysis. *Remote Sensing*, 14(4):940, 2022.

Larry J Sugarbaker, Eric W Constance, Hans Karl Heidemann, Allyson L Jason, Vicki Lukas, David L Saghy, and Jason M Stoker. The 3D elevation program initiative: a call for action. Technical report, US Geological Survey, 2014.

Towaki Takikawa, Thomas Müller, Merlin Nimier-David, Alex Evans, Sanja Fidler, Alec Jacobson, and Alexander Keller. Compact neural graphics primitives with learned hash probing, 2023. URL `https://arxiv.org/abs/2312.17241`.

Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, Börn Rommen, Nicolas Floury, Mike Brown, et al. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012.

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 & local features of many remote sensing modalities. In *Forty-second International Conference on Machine Learning*, 2025.

USDA. National Agriculture Imagery Program (NAIP). `https://naip-usdaonline.hub.arcgis.com/`, 2003–present. Accessed: November 23, 2025.

Marta Yebra, Gianluca Scortechini, Karine Adeline, Nursema Aktepe, Turkia Almoustafa, Avi Bar-Massada, María Eugenia Beget, Matthias Boer, Ross Bradstock, Tegan Brown, et al. Globe-LFMC 2.0, an enhanced and updated dataset for live fuel moisture content research. *Scientific Data*, 11(1):332, 2024.

APPENDICES

A EARTH4D RESOLUTION SPECIFICATIONS

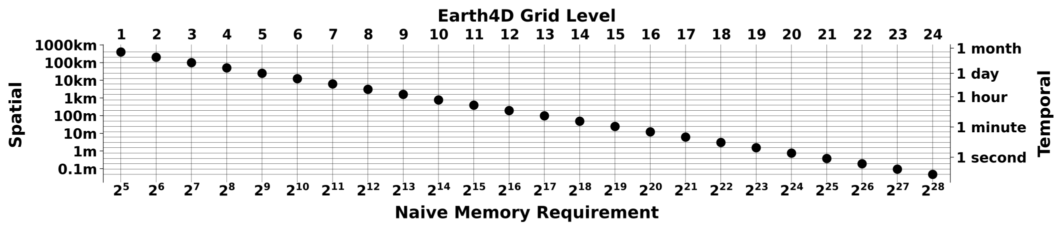

Figure 5: **Earth4D Space-Time Scales.** Default 24×24×24 levels for each *xyz*, *xyt*, *yzt*, *xzt* grid. Each level stores up to 2^{22} entries, with each entry storing a 2D feature. Requires 724M trainable parameters ($\sim$11 GB GPU memory during training). Parallelizable across levels and spatio-temporal boundaries. Outputs 192D per (x, y, z, t) coordinate from 4 grids $\times$ 24 levels $\times$ 2D feature per level. Hashing saves memory vs. naive requirement, e.g., $(2^{28})^3 = 10^{25}$ at level 24.

B Learned Hash Probing and Ablation Studies

B.1 Hash Collision Simulations

Scenario	Spatial	Temporal	Description
Uniform Random	Global	Full	Uniform Earth surface sampling
Continental Sparse	North America	Full	Sparse continental coverage
Moderate Spatial Cluster	10km × 10km	Full	City-scale clustering
Moderate Temporal Cluster	1k locations	Distributed	Temporal sampling at fixed locations
Moderate Spatiotemporal	1km × 1km	1 hour	Neighborhood-scale event
Extreme Spatial Single	10m × 10m	Full	Building-scale dense clustering
Extreme Spatial Multi	10 × (10m × 10m)	Full	10 dense clusters worldwide
Extreme Temporal Single	Global	1 hour	Global snapshot
Extreme Temporal Multi	Global	10 × (1 hour)	10 temporal snapshots
Time Series	10k locations	100 steps	Regular temporal sampling

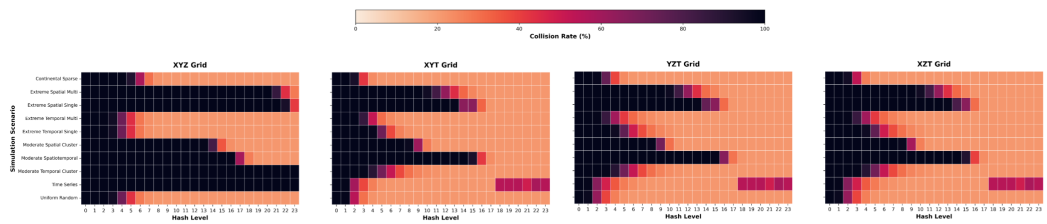

Figure 6: **Earth4D Hash Collision Analysis.** *(Table)* 10 (x, y, z, t) point distribution scenarios that were simulated to analyze hash collisions in Earth4D memory. *(Graph)* Shows results for 1M point simulations across all 24 levels.

B.2 Performance Improvements

Standard multi-resolution hash encoding without learned probing obtains RMSE 26.0pp, MAE 16.6pp, and R^2 0.58 (800M parameters, 2^{22} hash capacity). Integrating learned hash probing (Takikawa et al., 2023), which learns to select optimal hash table indices from a candidate set, yields RMSE 19.9pp, MAE 12.1pp, and R^2 0.755—a 27.1% MAE reduction and 30.2% R^2 improvement. Extreme compression to 5M parameters (99.3% reduction, 2^{14} hash capacity) achieves MAE 15.0pp/R^2 0.668, outperforming the 800M baseline by 14.7% in R^2 with 4× training speedup and 93% memory reduction. On RGB reconstruction, learned probing reduces validation loss by 18%. These gains result from collision reduction (33% at 1M points) and learned shared features across memory indices, allowing the model to discover meaningful spatio-temporal patterns.

C Benchmark Specifications

C.1 Galileo Baseline Model

Galileo (Johnson et al., 2025; Tseng et al., 2025) is a Vision Transformer (Dosovitskiy et al., 2021) (5.3M parameters) pre-trained by the Allen Institute for AI. It processes Sentinel-2 optical imagery (Drusch et al., 2012), Sentinel-1 SAR (Torres et al., 2012), VIIRS night lights, ERA-5 weather (Muñoz Sabater, 2019), TerraClimate soil/water data (Abatzoglou et al., 2018), SRTM topography (Farr & Kobrick, 2000), (x,y,z,t) coordinates, and species type. We use the Allen Institute for AI's exact Globe-LFMC 2.0 (Yebra et al., 2024) train/test split (76,467/13,297) to directly compare against this benchmark.