

Democratizing planetary-scale analysis: An ultra-lightweight Earth embedding database for accurate and flexible global land monitoring

Shuang Chen¹, Jie Wang^{2,*}, Shuai Yuan¹, Jiayang Li², Yu Xia², Yuanhong Liao³, Junbo Wei², Jincheng Yuan², Xiaoqing Xu², Xiaolin Zhu⁴, Peng Zhu¹, Hongsheng Zhang¹, Yuyu Zhou¹, Haohuan Fu⁵, Huabing Huang⁶, Bin Chen^{7,8}, Fan Dai⁸, Peng Gong^{1,8,9,*}

¹ Department of Geography, The University of Hong Kong, Hong Kong, China

² Pengcheng Laboratory, Shenzhen 518000, China

³ Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

⁴ Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China

⁵ Tsinghua Shenzhen International Graduate School, Shenzhen, China

⁶ School of Geospatial Engineering and Science, Sun Yat-sen University, Guangzhou 510275, China

⁷ Department of Architecture, The University of Hong Kong, Hong Kong, China.

⁸ Institute for Climate and Carbon Neutrality, The University of Hong Kong, Hong Kong, China

⁹ Department of Earth Sciences, The University of Hong Kong, Hong Kong, China

*Correspondence to: Jie Wang (wangj10@pcl.ac.cn) and Peng Gong (penggong@hku.hk)

Abstract. The rapid evolution of satellite-borne Earth Observation (EO) systems has fundamentally revolutionized terrestrial monitoring, yielding comprehensive petabyte-scale archives. However, the immense computational resources and storage volumes required for global-scale analysis often preclude widespread use by many research teams, hindering broader scientific adoption and the execution of planetary-scale studies. To address these barriers, we present the Embedded Seamless Data (ESD), an ultra-lightweight, 30-m global Earth embedding database spanning the 25-year period from 2000 to 2024.

By transforming high-dimensional, multi-sensor observations from the Landsat series (5, 7, 8, and 9) and MODIS Terra into information-dense, quantized latent vectors, ESD distills essential geophysical and semantic features into a unified latent space. Utilizing the ESDNet architecture and Finite Scalar Quantization (FSQ), the dataset achieves a transformative ~340-fold reduction in data volume compared to raw daily archives. This compression allows the entire global land surface for a single year to be encapsulated within approximately 2.4 TB, enabling decadal-scale global analysis on standard local workstations.

Rigorous validation demonstrates that ESD maintains high reconstructive fidelity to the original reflectance values across the spectral dimension, achieving a Mean Absolute Error (MAE) of 0.0130 (averaged over six spectral bands, including Blue, Green, Red, NIR, SWIR1, and SWIR2), a Root Mean Square Error (RMSE) of 0.0179, and a Correlation Coefficient (CC) of 0.8543. By condensing the annual phenological cycle into 12 temporal latent steps, the embeddings provide inherent denoising effects and a semantically organized latent space that outperforms raw reflectance data in downstream land-cover classification tasks, achieving a comparable and even higher overall accuracy of 79.74% than the 76.92% obtained using raw

sensor fusion data on globally distributed land cover sample sets. With robust few-shot learning capabilities and longitudinal consistency across 25 years, the ESD product provides a versatile foundation for democratizing planetary-scale Earth system research and advancing next-generation geospatial artificial intelligence.

1. Introduction

Over the past half-century, satellite-borne Earth Observation (EO) systems have fundamentally revolutionized our capacity to monitor terrestrial processes at a global scale (Drusch et al., 2012; Markham and Helder, 2012; Morisette et al., 2002; Wulder et al., 2022), yielding comprehensive datasets with unprecedented spatial and temporal continuity. These remotely sensed data have become pivotal to Earth system science, facilitating quantitative assessments of land surface dynamics (Brown et al., 2022; Gong et al., 2019, 2013; Liu et al., 2021; Zanaga et al., 2022), phenological cycles (Bolton et al., 2020; Piao et al., 2019), forest cover change (Aguirre-Gutiérrez et al., 2022; Forzieri et al., 2022; Hansen et al., 2013), inland water dynamics (Ji et al., 2018; Pekel et al., 2016; Pickens et al., 2022, 2020), and patterns of urban expansion (Gong et al., 2020, 2012; Huang et al., 2022).

Recent decades have seen the increasing availability of high-spatiotemporal-resolution satellite imagery and high-performance computing platforms, ushering in a new era for global-scale remote sensing applications (Drusch et al., 2012; Gorelick et al., 2017; Morisette et al., 2002; Wulder et al., 2022). For instance, coarse-spatial-resolution sensors, such as AVHRR and MODIS, have provided long-term records essential for climate variability and vegetation monitoring (Boyte et al., 2018; Chuvieco et al., 2005; Huete et al., 2002; Xu et al., 2022). Simultaneously, the medium-resolution Landsat missions have delivered the most extensive continuous record of land surface reflectance to date (Wulder et al., 2022). The adoption of a free and open data policy in 2008 (Woodcock et al., 2008) further accelerated the study of land-use change and ecosystem dynamics. More recently, the Sentinel-2 constellation has enhanced revisit frequencies, spatial resolutions, and spectral sampling, revealing new frontiers in precision agriculture and fine-scale land-cover mapping (Drusch et al., 2012).

Despite these advancements, many land monitoring applications necessitate EO data that possess both high spatial and high temporal resolutions, which remains unattainable from any single satellite sensor (Zhu et al., 2018). To bridge this gap, various multi-sensor fusion algorithms and datasets have been developed (Chen et al., 2024, 2023; Claverie et al., 2018; Gao et al., 2006; Goyena et al., 2023; Guo et al., 2020; Liu et al., 2022; Shang and Zhu, 2019; Zhu et al., 2016, 2010). These methods have enabled innovative applications in monitoring forest disturbances (Boyte et al., 2018; Hansen et al., 2008; Hilker et al., 2009; Shang et al., 2022) and surface water dynamics (Abowarda et al., 2021; Chen et al., 2018; Declaro and Kanae, 2024). Nevertheless, this rapid proliferation of sensor diversity and spatial-temporal-spectral resolution has resulted in an explosion of data volume, posing significant challenges for the storage, management, and processing of large-scale applications (Tuia et al., 2025).

While cloud-based platforms like Google Earth Engine and Microsoft Planetary Computer have democratized access to petabyte-scale datasets and scalable workflows (Gorelick et al., 2017; Smits, 2022; Zhao et al., 2021), significant barriers

remain. High computational costs, processing latencies, and associated financial constraints often preclude widespread use by smaller research teams, thereby hindering broader adoption and the execution of global-scale analyses (Tuia et al., 2025).

In parallel, artificial intelligence (specifically deep learning) has catalysed a paradigm shift in remote sensing data analysis (Tuia et al., 2025), achieving state-of-the-art performance in tasks such as land-cover classification (Ienco et al., 2019), semantic segmentation (Wieland et al., 2023), and change detection (Soto Vega et al., 2021). The recent advent of self-supervised pretraining (Devlin et al., 2019; He et al., 2022) and geospatial-specific foundation models has demonstrated remarkable generalization across multi-sensor and multi-modal EO inputs (Brown et al., 2025; Cong et al., 2022; Guo et al., 2024; Jakubik et al., 2023; Manas et al., 2021; Szwarcman et al., 2025). Nevertheless, the architectural complexity of these remote sensing foundation models (which frequently encompass billions of parameters) presents significant challenges. When the massive scale of these models is compounded by the vast data volumes inherent to global-scale remote sensing, the resulting computational costs for both pretraining and inference can become prohibitive. Consequently, despite their high performance and promise, these models remain difficult to deploy in many practical and operational settings due to the substantial resource requirements and the inherently immense size of remote sensing archives (Tuia et al., 2025).

Earth embeddings offer a promising solution by transforming high-dimensional, multi-modal satellite observations into compact, information-dense numerical vectors (Brown et al., 2025; Czerkawski et al., 2024; Feng et al., 2025; Klemmer et al., 2025). By projecting raw spectral data into a unified latent space, these embeddings distill essential geophysical and semantic features while significantly reducing data dimensionality (Brown et al., 2025). This approach not only mitigates the storage and processing bottlenecks inherent in petabyte-scale archives but also facilitates efficient downstream analysis, including cross-sensor fusion, classification, and change detection, without requiring the resource-intensive re-deployment of massive foundation models for every task (Brown et al., 2025; Klemmer et al., 2025).

Despite these theoretical advantages, there remains a notable scarcity of global-scale embedding datasets ready for immediate use by the broader scientific community (Klemmer et al., 2025). While pioneering initiatives such as Major TOM (Czerkawski et al., 2024), AlphaEarth Foundation (Brown et al., 2025), and TESSERA (Feng et al., 2025) represent a foundational precedent for the dissemination of large-scale geospatial embeddings, existing databases face critical limitations for operational Earth system monitoring (Klemmer et al., 2025). Specifically, next-generation embedding databases should provide: **(i) enhanced compression ratios** to further mitigate the storage and computational constraints of global-scale applications (Klemmer et al., 2025; Tuia et al., 2025); **(ii) extended temporal coverage** to support decadal-scale longitudinal studies; and **(iii) the explicit preservation of temporal structures** to characterize vital intra-annual dynamics and phenological variations (Klemmer et al., 2025).

To bridge these gaps, we present the Embedded Seamless Data (ESD), a global-scale 30-m Earth embedding dataset designed for accurate and efficient environmental monitoring. Derived from a fusion of Landsat-5/7/8/9 and MODIS Terra observations spanning from 2000 to 2024, ESD offers several transformative advantages over existing products. First, it achieves an exceptionally high compression ratio; the entire global land surface for a single year is encapsulated within approximately 2.4 TB, representing a significant reduction in storage requirements that democratizes petabyte-scale analysis

for researchers with limited computational infrastructure. Second, ESD explicitly preserves temporal structures and information, providing the latent depth necessary to resolve intra-annual dynamics and complex phenological cycles. Third, the high fidelity of these embeddings allows for the reconstruction of near-daily 30-m observations with remarkably low error rates, providing a continuous proxy for surface processes. Finally, ESD provides extended temporal coverage spanning from 2000 to 2024, offering the longitudinal consistency required to track decadal shifts in Earth system processes.

The remainder of this paper is organized as follows: **Section 2** describes the multi-source satellite observations used to develop and validate the ESD dataset. **Section 3** outlines the methodological framework for the construction of the embeddings and the protocols used for accuracy assessment. In **Section 4**, we present the comprehensive validation results alongside several case studies that demonstrate the dataset’s utility in monitoring land surface dynamics. **Section 5** discusses the technical advantages of ESD, its inherent limitations, and its potential for future Earth system research. Finally, **Section 6** provides details on data availability and access through open-science repositories, followed by concluding remarks in **Section 7**.

2. Materials

2.1 Input satellite imagery and auxiliary dataset

2.1.1 Global 30m Seamless Data Cube of Land Surface Reflectance

The development of the Embedded Seamless Data (ESD) is founded upon the integration of two primary satellite constellations: the Landsat series (5, 7, 8, and 9) and MODIS (Terra). As listed in **Table 1**, we utilize six spectral bands: blue, green, red, near-infrared (NIR), and two shortwave infrared (SWIR1 and SWIR2), which provide the necessary spectral range for comprehensive environmental monitoring and land-surface characterization.

To generate a gap-free foundation for our embeddings, we first synthesized these multi-sensor observations into the Global 30 m Seamless Data Cube (SDC30) of Land Surface Reflectance (Chen et al., 2024). This was achieved through a rigorous processing pipeline encompassing data harmonization, missing-value reconstruction, and spatiotemporal fusion. The resulting SDC30 provides a spatially consistent, 30 m resolution record of global surface reflectance spanning the 25-year period from 2000 to 2024. This seamless data cube serves as the direct input for the construction of the ESD, transforming the reconstructed reflectance values into high-efficiency, information-dense latent representations.

Table 1: Attributes of the six employed spectral bands from Landsat 5 TM, 7 ETM+, 8-9 OLI, and MODIS Terra products (Markham and Helder, 2012; Masek et al., 2020; Morisette et al., 2002).

Wavelengths (micrometers)					
	Landsat 5 TM	Landsat 7 ETM+	Landsat 8 OLI	Landsat 9 OLI-2	MODIS Terra

Blue	0.450-0.520	0.450-0.515	0.452-0.512	0.452-0.512	0.459-0.479
Green	0.520-0.600	0.525-0.600	0.533-0.590	0.532-0.589	0.545-0.565
Red	0.630-0.690	0.630-0.690	0.636-0.673	0.636-0.672	0.620-0.670
NIR	0.760-0.900	0.760-0.900	0.851-0.879	0.850-0.879	0.841-0.876
SWIR1	1.550-1.750	1.550-1.750	1.566-1.651	1.565-1.651	1.628-1.652
SWIR2	2.080-2.350	2.080-2.350	2.107-2.294	2.105-2.294	2.105-2.155
Spatial resolution	30-m	30-m	30-m	30-m	500-m (near)
Revisit frequency	16-day	16-day	16-day	16-day	Daily (near)

2.1.2 NASADEM: NASA 30m Digital Elevation Model

To incorporate topographic context into the embeddings, we utilized the NASADEM Global Digital Elevation Model (GDEM) at a 30 m spatial resolution (NASA JPL, 2020). NASADEM represents a major refinement of the Shuttle Radar Topography Mission (SRTM) archive, achieving superior vertical accuracy and data continuity by synthesizing SRTM data with auxiliary high-quality elevation sources. Specifically, it integrates observations from the ASTER GDEM, the ICESat Geoscience Laser Altimeter System (GLAS), and the ALOS Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) to eliminate voids and improve the representation of complex terrain. Within the ESD framework, this elevation layer provides essential terrain-dependent features (such as slope, aspect, and elevation) that act as critical static covariates. These features enhance the model's capacity to characterize land-surface processes and improve the performance of downstream applications in hydrological modeling, ecosystem mapping, and terrain-aware change detection.

2.2 Supervisory land cover datasets

To ensure that the learned latent representations within the ESD align with meaningful biophysical properties and high-level semantic categories, we incorporated several authoritative land cover products as supervisory signals during the model training phase. By constraining the embedding space with these thematic labels, the model is better equipped to distill essential geophysical features from raw reflectance, thereby enhancing the interpretability and performance of the embeddings in downstream classification and change-detection tasks.

2.2.1 Annual maps of global artificial impervious area

The Global Artificial Impervious Area (GAIA) dataset (Gong et al., 2020) was utilized to provide long-term supervision for built-up environments. Derived from the Landsat archive at a 30 m spatial resolution, GAIA provides annual global maps of impervious surfaces from 1985 to 2024. This dataset is characterized by high thematic accuracy (exceeding 90%) and is particularly effective in mitigating the spectral confusion between building shadows and open water bodies, a common

challenge in urban remote sensing. In this study, GAIA serves as a temporal anchor to ensure the stability of the embeddings in rapidly urbanizing regions over the 25-year study period.

2.2.2 ESA WorldCover 2021

The ESA WorldCover 2021 dataset (Zanaga et al., 2022), a global land cover product at 10-meter spatial resolution, was employed as supervisory data in the training phase of this study. Generated from Sentinel-1 and Sentinel-2 satellite imagery, this dataset provides detailed and reliable land cover classifications encompassing multiple land cover categories, including forests, urban areas, croplands, water bodies, and various natural vegetation types. The high-resolution and globally consistent nature of ESA WorldCover 2021 made it an ideal supervisory dataset, significantly enhancing the semantic quality and interpretability of the generated embedding vectors.

2.2.3 GLAD global crop extent dynamics

The GLAD Global Crop Extent Dynamics (GLAD-CE) dataset (Potapov et al., 2021) was employed to supervise the identification of agricultural landscapes. This 30 m product provides a consistent time series of cropland extent from 2000 to 2019, partitioned into four-year intervals. Following the GLAD definition, cropland includes land used for annual and perennial herbaceous crops while excluding woody perennials and permanent pastures. The inclusion of GLAD-CE ensures that the resulting ESD embeddings accurately preserve the phenological signatures unique to intensive agriculture and crop rotation cycles (<https://glad.umd.edu/dataset/croplands>; last accessed: December 2, 2024).

2.2.4 GLAD global monthly surface water dynamics

To characterize intra-annual hydrological variations, we leveraged the Global Land Analysis and Discovery Global Surface Water Dynamics (GLAD-SW) dataset (Pickens et al., 2022, 2020). This product provides monthly global surface water extent at a 30 m resolution from 1999 to 2021. With reported user's and producer's accuracies of 93.7% and 96.0%, respectively, GLAD-SW provides the necessary temporal supervision to resolve seasonal and inter-annual fluctuations in aquatic ecosystems. The dataset was accessed via the official GLAD repository (<https://glad.umd.edu/dataset/global-surface-water-dynamics>; last accessed: December 2, 2024).

2.3 Validation datasets

2.3.1 FAST: First All-season Sample seTs for global land cover classification

The First All-Season Sample Set (FAST) serves as the primary reference for benchmarking our classification performance. FAST is an evolution of the Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) project (Gong et al., 2013), comprising 91,619 training and 36,636 validation locations worldwide.

185 The spatial allocation of validation samples followed a rigorous hex-grid approach, where the global land surface was partitioned into approximately 7,000 equal-area hexagons, with five pixel locations randomly selected within each. To capture phenological variability, the original single-date samples were expanded to a seasonal framework using primarily Landsat 8 OLI surface reflectance from 2014–2015 (Li et al., 2017). This seasonal expansion was manually verified by a team of expert interpreters using a multi-source evidence approach, integrating MODIS EVI time series, monthly climatology (temperature and precipitation), and high-resolution imagery from Google Earth.

190 The FAST classification system encompasses 11 primary categories: cropland, forest, grassland, shrubland, wetland, water, tundra, impervious surface, bareland, snow/ice, and cloud. For the purposes of this study, we aggregated seasonal labels into annual representations and excluded cloud-contaminated samples. To ensure spatial precision in the 30 m embedding process, only the central pixel of each sample unit was utilized for training and validation.

3. Methodology

195 3.1 Overview of the ESD production pipeline

This section outlines the integrated framework developed to generate the Embedded Seamless Data (ESD), transitioning from raw multi-sensor observations to information-dense latent representations. As illustrated in **Figure 1**, the pipelined ESDNet is organized into three primary modules:

- (a) **Spatiotemporal Fusion:** The harmonization of heterogeneous Landsat and MODIS archives into the unified 30 m Seamless Data Cube (SDC30) (Chen et al., 2024).
- (b) **Latent Encoding and Quantization:** A deep learning-based encoding process that transforms high-dimensional spectral-temporal features into quantized embedding vectors.
- (c) **Multitask Inference and Reconstruction:** A suite of specialized task heads $\{D_1, D_2, D_3, D_4, \dots\}$ that leverage the embeddings to generate downstream information products and reconstruct the original SDC30 inputs.

205

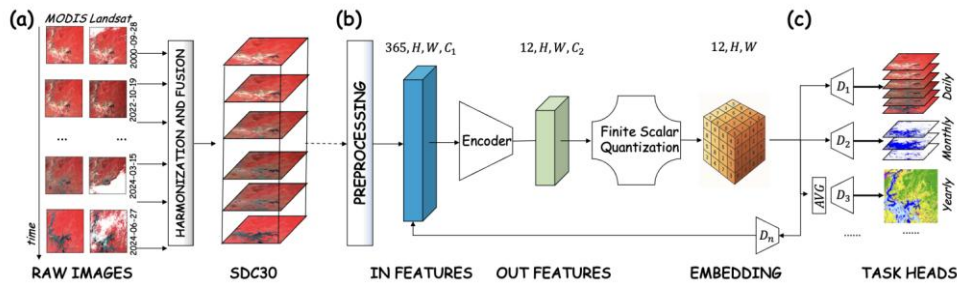


Figure 1: Schematic of the integrated ESD production pipeline. (a) Heterogeneous Landsat and MODIS time-series observations are harmonized and fused into the unified Global 30-m Seamless Data Cube (SDC30). (b) For each processing tile, the annual SDC30 feature cube of shape $[365, H, W, C_1]$ is preprocessed and then transformed by the Encoder into a compact latent representation, which is then discretized by the Finite Scalar Quantization (FSQ) module into 12 monthly latent steps of shape $[12, H, W, C_2]$. (c) The quantized embeddings are used by the Decoder to reconstruct the original surface-reflectance time series and by the MLP task heads to impose auxiliary thematic supervision during training, yielding the released ESD product. Here, 365 denotes the annual daily axis of the SDC30 cube. H and W denote the height and width of a processing tile. C_1 denotes the per-time-step input feature dimensionality after preprocessing. The internal latent representation contains 12 monthly steps, and C_2 denotes the latent dimensionality at each step. The distributed annual ESD product retains the 12-step temporal axis and stores one quantized token index per step, yielding a tensor of shape $[12, H, W]$. In the released configuration, $C_2 = 6$.

批注 [JL1]: Figure caption revised

3.2 ESDNet Architecture and Computational Framework

The ESDNet architecture is engineered to transform time-series spatiotemporal reflectance data into discrete, information-dense embeddings. High-frequency noise is mitigated jointly by the upstream SDC30 harmonization process and by the architecture of ESDNet itself. Before entering the network, Landsat and MODIS observations are harmonized through radiometric normalization, temporal gap filling, and cloud/shadow reconstruction, which reduce short-lived contamination and sensor-dependent artifacts. Within ESDNet, the strided Conv1D layers aggregate information over longer temporal windows, and the FSQ bottleneck further suppresses weak perturbations by discretizing each latent dimension onto a fixed scalar lattice. As illustrated in Figure 2, the network employs a symmetric encoder-decoder structure integrated with a discrete quantization bottleneck and multi-task prediction heads.

批注 [JL2]:

(a) Temporal Encoder Network

The encoder is designed to extract hierarchical features from the SDC30 time-series. It consists of N initial 1D convolutional (Conv1D) layers followed by M residual Conv1D blocks. To reduce computational overhead and condense the temporal signal, the initial layers utilize a stride $s > 1$, performing temporal downsampling. This effectively reduces the input sequence length while broadening the receptive field to capture long-term phenological patterns. The subsequent residual blocks employ a stride of 1 and incorporate shortcut connections to facilitate the training of deeper configurations, mitigating gradient vanishing issues and ensuring the preservation of high-level semantic abstractions.

(b) Finite Scalar Quantization (FSQ) Module

To achieve the high compression ratios required for a global-scale database, the latent features are passed through a Finite Scalar Quantization (FSQ) module (Mentzer et al., 2023). Unlike traditional Vector Quantization (Oord et al., 2018), which relies on a learned codebook prone to "codebook collapse," FSQ projects latent representations into a bounded, lower-dimensional space where each dimension is quantized into a set of fixed, predefined scalars.

The quantization is implemented via a simple rounding operation, and gradients are propagated during backpropagation using a straight-through estimator (STE) (Bengio et al., 2013). This design creates an implicit codebook via the Cartesian product of the scalar sets, providing a highly stable and lightweight quantization scheme that eliminates the need for auxiliary commitment losses (Takida et al., 2022).

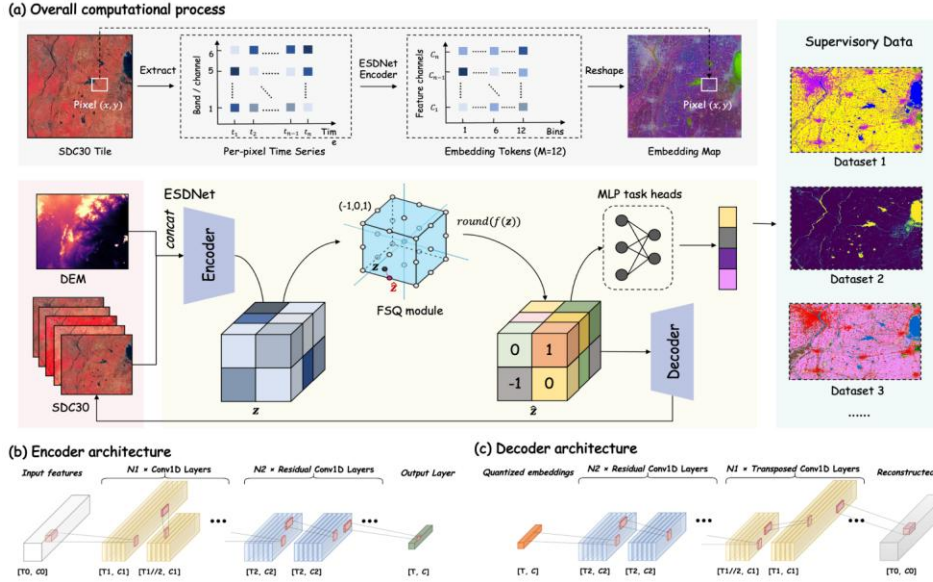


Figure 2: Detailed network architecture and computational process of ESDNet. (a) Overall workflow. For each pixel (x,y) in an SDC30 tile, the annual input feature sequence is passed to the Encoder, which compresses the signal into $M=12$ monthly tokens. These tokens are discretized by the Finite Scalar Quantization (FSQ) module, which maps the encoder output onto a fixed scalar lattice within a bounded latent space. The quantized monthly tokens are then routed to two branches. The first branch is the Decoder, which reconstructs the input temporal sequence and preserves the spectral-temporal information required for reconstructive fidelity. The second branch consists of the MLP task heads, which generate supervised predictions aligned with the auxiliary thematic products used in training. (b) Encoder architecture. The Encoder begins with N_1 strided Conv1D layers that progressively shorten the temporal dimension and enlarge the effective temporal receptive field, followed by N_2 residual Conv1D layers for deeper feature extraction, and a final pointwise projection that maps the hidden representation into the latent embedding space before quantization. (c) Decoder architecture. The Decoder mirrors the Encoder by first expanding the latent representation through residual Conv1D layers and then progressively restoring the original temporal length through transposed Conv1D upsampling layers. Tensor shapes are annotated at each stage as $[T, C]$, where T denotes temporal length and C denotes channel width. Here, T_0 denotes the annual input sequence length, $M=12$ denotes the monthly tokens at the bottleneck, and C_0 , C_1 , and C_2 denote the channel widths at the input, intermediate convolutional, and residual stages, respectively.

批注 [JL3]: Figure caption revised

(c) Mirror Decoder and Reconstruction

The decoder serves as the architectural mirror of the encoder, comprising M residual blocks and N transposed Conv1D layers. The transposed convolutions upsample the quantized embeddings, progressively restoring the original temporal dimensions of the input SDC30 reflectance data. The primary objective of the decoder is the high-fidelity reconstruction of the input sequence; the resulting reconstruction loss provides a self-supervised signal that ensures the embeddings retain the fundamental biophysical properties of the land surface.

265

(d) Multi-Task Prediction Heads

270

To imbue the embeddings with explicit semantic meaning, the quantized latent vectors are processed through a temporal average-pooling layer to generate a summary feature vector. This vector is fed into multiple, independent task-specific heads, each implemented as a single-layer fully connected network. These heads are trained concurrently using the supervisory datasets described in **Section 2.2** (e.g., GAIA, ESA WorldCover, GLAD-SW, and GLAD-CE). This multi-task learning (MTL) framework constrains the latent space, forcing the embeddings to represent both spectral-temporal dynamics and high-level land-system categories.

275

[The released ESDNet configuration uses three strided Conv1D layers in the Encoder to progressively reduce the temporal dimension, followed by a stack of 10 residual blocks and a pointwise projection into a 6-dimensional latent space. The FSQ bottleneck produces 12 monthly tokens per pixel-year. The Decoder mirrors this structure with residual blocks followed by three transposed Conv1D layers that restore the temporal dimension and reconstruct the six surface-reflectance bands. The main architectural hyperparameters of the released configuration are summarized in **Table 2**. Specifically, for the released compression factor of 8, the Encoder applies three Conv1D layers with kernel size 4, stride 2, and padding 1, followed by a Conv1D layer with kernel size 3 and stride 1 before the residual stack. Each residual block contains a Conv1D layer with kernel size 3 and a pointwise Conv1D layer with kernel size 1. The Decoder mirrors this design, beginning with a Conv1D layer with kernel size 3 and stride 1 and then using three transposed Conv1D layers with kernel size 4, stride 2, and padding 1 to restore the temporal dimension.]

280

批注 [JL4]:

Table 2: Architectural hyperparameters of the released ESDNet configuration.

批注 [JL5]: Hyperparameters added

Hyperparameter	Value
Input feature dimensionality	13
Reconstruction target dimensionality	6
Hidden channel width C	32
Residual block hidden width	256
Number of stride-2 sampling stages N_1	3
Number of residual blocks N_2	10
Temporal compression factor	8
Embedding dimensionality d	6
Tokens per pixel-year M	12

285

3.3 Multi-Task Training Strategy

To ensure that the Embedded Seamless Data (ESD) captures both the fundamental biophysical properties and the high-level semantic categories of the land surface, we implemented a multi-task learning (MTL) framework. The model is trained using

a joint objective function that combines an unsupervised reconstruction task with supervised classification and regression tasks. The total loss L is defined as a weighted sum of these components:

$$L = \alpha L_{reconstruction} + \beta L_{classification} + \gamma L_{regression} \quad (1)$$

where α , β , and γ are hyperparameters that balance the contribution of each task to the total gradient. For the released configuration, α was fixed at 1.0 and $\beta = \gamma$ were jointly set to 0.1 as an empirically balanced weighting, so that reconstruction remained the dominant objective while the auxiliary supervision provided a weaker constraint to improve thematic separability. The three loss terms in Eq. (1) are optimized jointly rather than in separate stages. At each training step, the input mini-batch is forwarded through the Encoder and the FSQ bottleneck to produce the quantized latent representation. The Decoder computes the reconstruction term, while the parallel supervised branches compute the auxiliary supervision terms. These components are combined into a single scalar objective according to Eq. (1), and a single backward pass updates all trainable parameters simultaneously.

批注 [JL6]:

(a) Unsupervised Reconstruction Loss

The reconstruction loss, $L_{reconstruction}$, quantifies the discrepancy between the input reflectance sequence x and the reconstructed output $\hat{x}$ generated by the decoder (e.g., D_n in **Figure 1**). Following a Gaussian likelihood assumption for pixel-level values, this is implemented as a Mean Squared Error (MSE) objective:

$$L_{reconstruction} = \log p(x|\hat{z}(x)) = \|x - \hat{x}\|^2 \quad (2)$$

This self-supervised signal ensures that the quantized embeddings preserve the fine-scale spectral signatures and phenological variations necessary for near-daily and monthly surface monitoring.

(b) Supervised Classification Loss

The classification loss $L_{classification}$ is computed across multiple task-specific heads (e.g., D_1 , D_2 , and D_3 in **Figure 1**). These heads are supervised by the diverse land-cover products described in **Section 2.2** (e.g., FROMGLC, GAIA and ESA WorldCover). For each supervisory task i , the loss is calculated using weighted multi-label cross-entropy:

$$L_{classification} = \sum_i a_i \sum_{k=1}^{K_i} y_k \log(\hat{y}_k) \quad (3)$$

where y_k represents the ground-truth label, $\hat{y}_k$ is the predicted probability for class k , K_i denotes the number of classes in supervisory task i , and a_i is a task-specific weight.

批注 [JL8]:

This constraint forces the latent space to cluster according to meaningful thematic categories, such as forest, water, or impervious surfaces.

(c) Supervised Regression Loss

To further align the embeddings with essential climate and vegetation variables, we incorporate a regression objective, $L_{regression}$. This task focuses on predicting biophysical indices derived from the SDC time series and aggregated as monthly averages, such as the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Snow Index (NDSI). The loss is formulated as:

批注 [JL9]:

320

$$L_{regression} = \sum_i b_i \|v - \hat{v}\|^2 \tag{4}$$

where v and $\hat{v}$ are the target and predicted biophysical indices, respectively, and b_i is a per-variable weight. By explicitly regressing these indices, the model learns to prioritize the spectral bands and temporal windows most critical for characterizing vegetation health and water dynamics.

批注 [JL10]:

3.4 Global-scale Model Training and Validation

325 3.4.1 Training Samples Generation and Harmonization

To ensure robust model generalization across diverse geographic regions, climatic zones, and ecological biomes, we constructed a large-scale, globally distributed training dataset. We selected a total of 223,622 unique 6×6 km locations through a rigorous stratified random sampling strategy. To ensure the dataset represents the full spectrum of Earth’s surface characteristics, stratification was guided by three primary dimensions: biome boundaries, continental divisions, and land-cover class proportions derived from the ESA WorldCover 2021 global map. This approach ensures adequate representation of all major thematic classes, including forests, shrublands, grasslands, croplands, urban areas, barren lands, wetlands, water bodies, and snow/ice, across varying environmental and socio-ecological contexts.

330

For each sampling unit, we synthesized a comprehensive suite of multi-temporal and multi-source features spanning the period from 2000 to 2024 (Table 3). The model’s input features are categorized into two streams:

335

- **Dynamic Inputs:** Annual SDC30 time-series surface reflectance (30 m resolution), providing the primary spectral-temporal signal. The input shape of $[365 \times 6]$ represents the daily observations across six spectral bands, capturing the full annual phenological cycle.
- **Static Covariates:** Topographic attributes derived from NASADEM (e.g., elevation and slope), providing spatially informative context to support terrain-aware feature extraction.

340

To reconcile the heterogeneous temporal cadences of the supervisory labels, all products were resampled to a common 30 m grid and synchronized with the SDC30 reflectance records. As detailed in Table 3, we adopted a tiered synchronization logic: static products (e.g., NASADEM, WorldCover 2021) were treated as invariant across the study period; annual products (FROMGLC30, GAIA) were matched by calendar year; and composite products (e.g., GLAD-CE) were assigned to their corresponding four-year windows. For the GLAD-SW dataset, we maintained its full monthly granularity by linking each annual SDC30 observation to twelve monthly water-presence indicators, enabling the model to better characterize intra-annual hydrological variability and seasonality.

345

Table 3: Attributes of the multi-source satellite imagery and supervisory datasets used for ESDNet training.

Product	SDC30	NASADEM	WorldCover2021	FROMGLC30	GAIA	GLAD-CE	GLAD-SW
Cadence	Yearly	Static	Static	Yearly	Yearly	4-Year	Monthly

Coverage	2000-2024	Static	2021	2000-2024	1985-2024	2000-2019	1999-2024
Data shape	[365×6]	[1]	[1]	[1]	[1]	[1]	[12]

The spatial distribution of these sampled locations is illustrated in **Figure 3**, colored by their dominant land-cover class. This highly diverse and balanced training foundation ensures broad spectral and spatial coverage, providing the statistical depth necessary for accurate, long-term monitoring of global land-cover dynamics. For reproducibility, the optimization and deployment settings of the released configuration are summarized as follows. Training used the Adam optimizer with a learning rate of 5×10^{-4} , without warmup or learning-rate decay. The public training script defaults to a batch size of 1024, uses 8 data-loading workers, evaluates every 2000 steps, saves checkpoints every 10000 steps, and trains for up to 1000 epochs with early stopping after 10 evaluation intervals without improvement. No custom weight-initialization scheme was applied beyond the default PyTorch initialization. For the released loss configuration, $\alpha=1.0$ and $\beta=\gamma=0.1$. For inference, annual production tiles are generated tile-wise at 3600×3600 pixels and partitioned into non-overlapping 600 × 600 subwindows for memory-efficient processing. Each annual sequence is converted into 13 input features and temporally interpolated to length 96 before encoding. No overlap blending is used during tile generation.

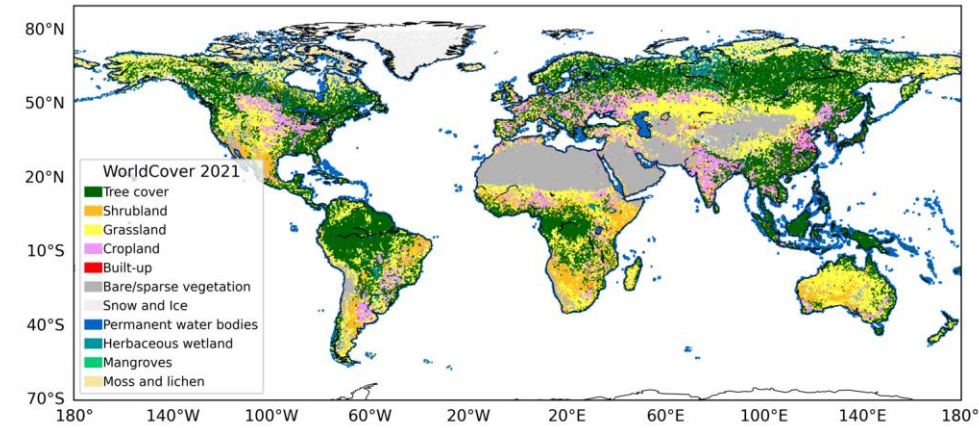


Figure 3. Geographic distribution and thematic composition of the global training dataset. The map displays the locations of the 223,622 stratified random sampling sites used to train the ESDNet. These sites were generated using a spatially uniform global sampling design, so their high density may visually resemble a wall-to-wall land-cover product in some regions. Each site is colored according to its dominant land-cover class as derived from the ESA WorldCover 2021 product.

3.4.2 Quantitative Evaluation: Reconstruction Fidelity of the Discrete Latent Space

To evaluate the representational capacity of the learned embeddings, we assessed the reconstructive fidelity of the ESDNet encoder-decoder architecture. This evaluation quantifies the degree to which essential spectral and temporal information is preserved within the compressed, quantized latent space. Specifically, we measured the discrepancy between the original SDC30 input reflectance (x) and the corresponding reconstructed output ($\hat{x}$) generated by the decoder.

Reconstruction accuracy was primarily quantified using the Mean Absolute Error (MAE), calculated independently for each of the six spectral bands across the full 365-day temporal sequence. To provide a comprehensive assessment of physical fidelity, we also computed the Root Mean Square Error (RMSE) and the Correlation Coefficient (CC). The evaluation was conducted using the FAST-validation repository, which consists of 36,636 globally distributed locations. These samples were strictly held out during the training phase to ensure an unbiased assessment of the model's generalization capabilities across diverse terrestrial biomes and phenological regimes.

3.4.3 Evaluation of Generalization and Transferability to Unseen Tasks

A core objective of the ESDNet is to provide "universal" representations that support generalization beyond the specific supervised tasks encountered during training. To test this capability, we conducted two transfer learning experiments where the learned embeddings were utilized as fixed, frozen feature representations for novel classification problems.

In-Domain Transfer: We trained a diverse set of downstream classifiers, including linear prediction heads, Ridge Classifiers, K-Nearest Neighbors (KNN), and Random Forests, using the FAST-training set and evaluated their performance on the FAST-validation set.

Out-of-Domain Generalization: To assess the model's robustness against external benchmarks, we utilized the official training and validation samples provided by the Copernicus Global Land Operations Service (CGLOPS).

For both experiments, we employed the same suite of performance metrics (OA, precision, recall, and F1-score). Furthermore, we analyzed model performance under varying data regimes to evaluate the few-shot learning capabilities of the ESD embeddings. By measuring accuracy across a range of training sample sizes, this analysis provides critical insights into the practical utility of the representations in data-scarce settings and their adaptability to emergent downstream applications in Earth system science.

4. Results and Analysis

4.1 Global Product Characteristics and Data Structure

The Embedded Seamless Data (ESD) provides a high-fidelity, continuous 30 m resolution representation of the global land surface, spanning a 25-year longitudinal record from 2000 to 2024. By leveraging the robust reconstructive capabilities of the ESDNet decoder and the initial spatiotemporal fusion of the SDC30 foundation, the resulting embeddings effectively

eliminate the traditional barriers to global-scale analysis, such as "seam-line" artifacts, cloud-contamination, and sensor-specific biases.

As illustrated in **Figure 4**, the ESD maintains remarkable spatial and spectral consistency across continental scales. Despite the high compression ratio, the embeddings preserve fine-grained landscape features, including intricate river networks, urban morphological boundaries, and fragmented agricultural patterns, ensuring that the latent space acts as a reliable proxy for raw surface reflectance.

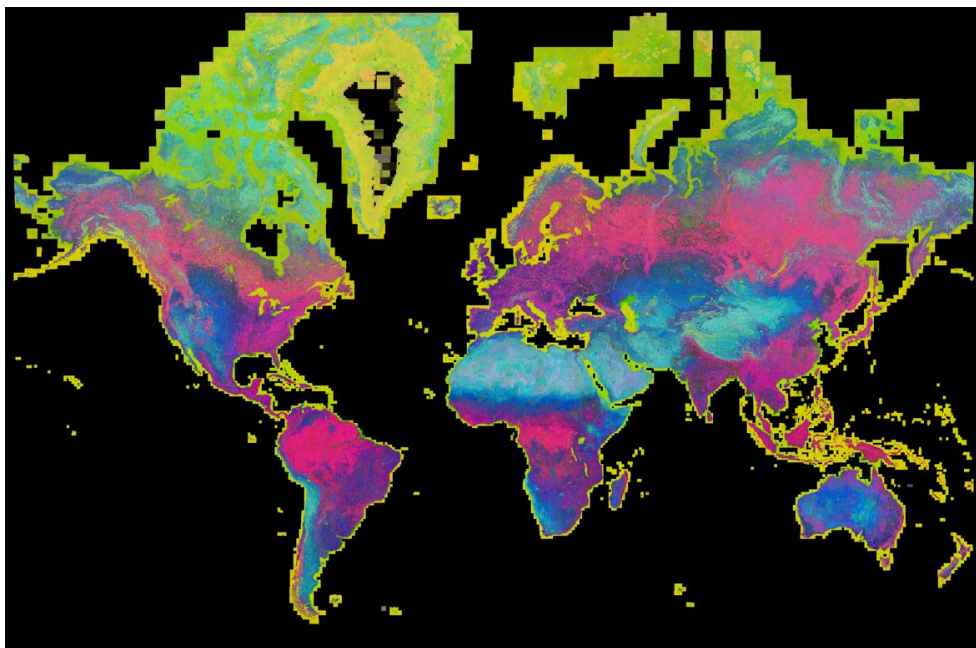


Figure 4: Global visualization of the 2024 ESD product at 30-m resolution. This false-color composite represents three selected dimensions of the information-dense latent vectors.

To ensure maximum interoperability with existing Earth Observation (EO) workflows, the ESD is organized according to the Military Grid Reference System (MGRS) in the Universal Transverse Mercator (UTM) projection, matching the tiling convention of the Sentinel-2 mission (Drusch et al., 2012).

- **Dimensions:** Every tile follows a fixed structure of [12, 3600, 3600]. The first dimension (12) represents 12 temporal latent steps, which correspond to monthly information-dense summaries that capture intra-annual

the NIR band exhibited the highest absolute error (MAE = 0.0170). This variation is consistent with the higher dynamic range and inherent phenological variability typically found in near-infrared observations of vegetated surfaces. The mean Correlation Coefficient (CC) remains high at 0.854. This confirms that the Finite Scalar Quantization (FSQ) bottleneck effectively retains the spectral-temporal signatures necessary for resolving fine-grained land surface dynamics. A closer inspection of Table 5 reveals systematic differences in reconstructive performance across the six spectral bands that reflect their underlying physical and statistical properties. The visible bands (Blue, Green, and Red) show consistently low MAE values and moderate-to-high CC values, which is consistent with their relatively narrow dynamic range over typical land surfaces. The NIR band exhibits the largest MAE but also the highest CC, indicating that although the absolute reconstruction error increases with its much larger reflectance range, the seasonal trajectory is still reproduced faithfully. SWIR1 shows intermediate behavior, while SWIR2 combines the lowest MAE with a comparatively lower CC, which we attribute to its generally low and spatially stable reflectance levels: these reduce absolute error but also provide less variance for the embedding to match.

Table 5: Decoder reconstruction accuracy in MAE, RMSE, and CC evaluated using the FAST-validation set.

Band	Blue	Green	Red	NIR	SWIR1	SWIR2	Mean
MAE↓	0.0121	0.0116	0.0123	0.0170	0.0139	0.0115	0.0130
RMSE↓	0.0176	0.0166	0.0175	0.0227	0.0183	0.0150	0.0179
CC↑	0.8023	0.8494	0.8801	0.8848	0.8681	0.8411	0.8543

A critical requirement for Earth System Science Data is the stability of the signal over time. Temporal analysis illustrated in Figure 5 reveals that the reconstructive fidelity is remarkably consistent throughout the 25-year study period.

The stability of the MAE and CC metrics over two decades demonstrates the model's robustness against transitions between underlying satellite constellations, specifically from Landsat-5 TM and 7 ETM+ to the newer Landsat-8/9 OLI sensors. This longitudinal consistency ensures that the ESD can be used for tracking decadal shifts in Earth system processes, such as forest degradation, urban expansion, and water body fluctuations, without introducing artificial "jumps" caused by sensor changes.

批注 [JL13]:

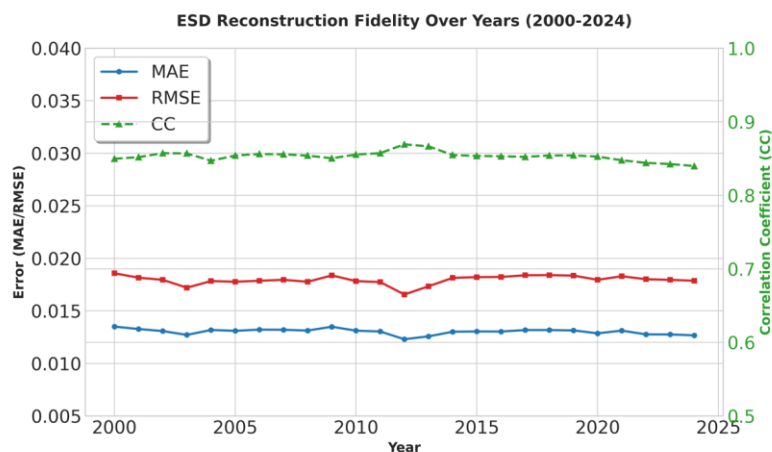


Figure 5: Temporal evaluation of the reconstructive fidelity of the ESDNet encoder-decoder architecture (2000–2024). The plot illustrates the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Correlation Coefficient (CC) calculated between the input SDC30 surface reflectance and the reconstructed output generated from the quantized latent space. Metrics represent the average performance across all six spectral bands (Blue, Green, Red, NIR, SWIR1, and SWIR2). Validation was performed using the globally distributed FAST-validation repository, comprising independent samples held out during the model training phase.

Beyond statistical metrics, the ESDNet's ability to preserve the structural integrity of the landscape is vital. As shown in the detailed tile evaluations (Figures 7–10), the model effectively reconstructs essential landscape features even in regions with challenging conditions: a) Intricate river networks, crisp urban boundaries, and individual agricultural field patterns are clearly maintained in the reconstructed output; b) The reconstruction process demonstrates a notable "denoising" effect, where residual artifacts from cloud-contamination or topography in the original SDC30 data are minimized in the latent representation; c) The spatial distribution of MAE, visualized in Figure 6, confirms that low error rates are achieved globally, across diverse terrestrial biomes and extreme topographic reliefs.

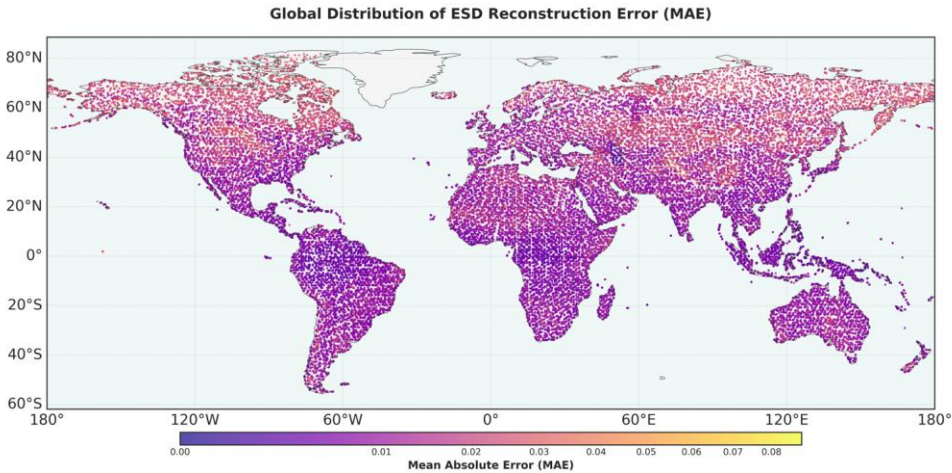


Figure 6: Global spatial distribution of reconstructive fidelity in MAE for the Embedded Seamless Data (ESD) product.

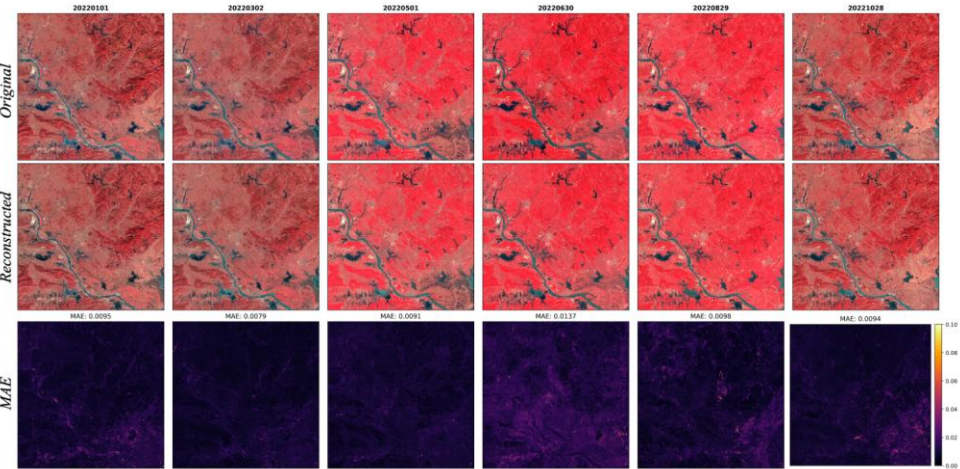


Figure 7: Reconstructive performance in tile 50RLU. The top row displays the original SDC30 surface reflectance (NIR-Red-Green false-color composite). The middle row shows the corresponding reconstructed output from the ESDNet decoder. The bottom row quantifies the spatial distribution of the MAE per pixel.

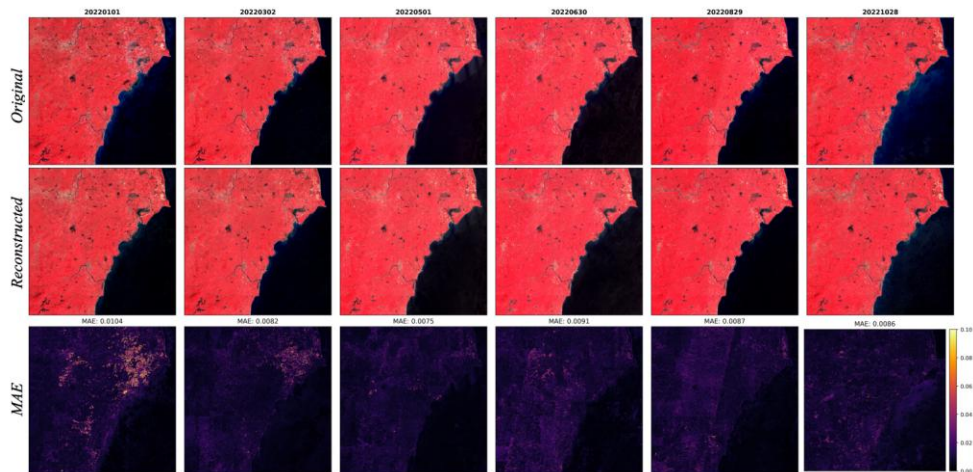


Figure 8: Reconstructive performance in tile 49QDB. The top row displays the original SDC30 surface reflectance (NIR-Red-Green false-color composite). The middle row shows the corresponding reconstructed output from the ESDNet decoder. The bottom row quantifies the spatial distribution of the MAE per pixel.

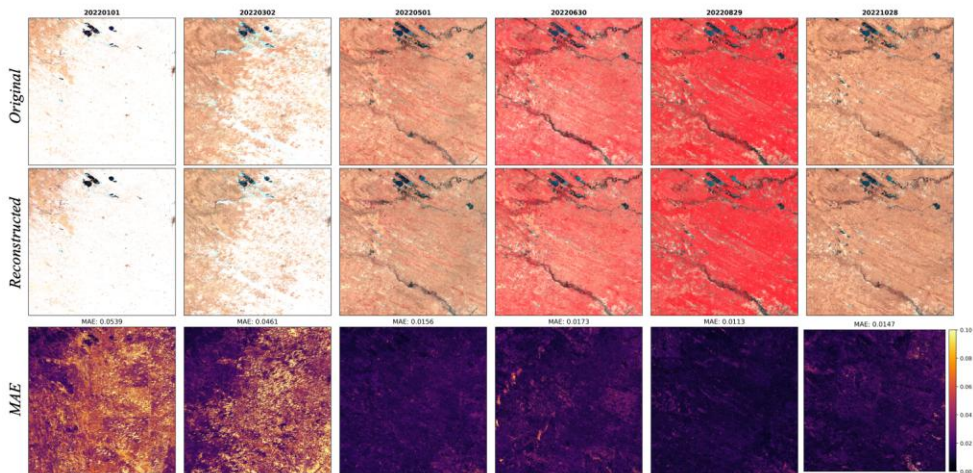


Figure 9: Reconstructive performance in tile 51TVK. The top row displays the original SDC30 surface reflectance (NIR-Red-Green false-color composite). The middle row shows the corresponding reconstructed output from the ESDNet decoder. The bottom row quantifies the spatial distribution of the MAE per pixel.

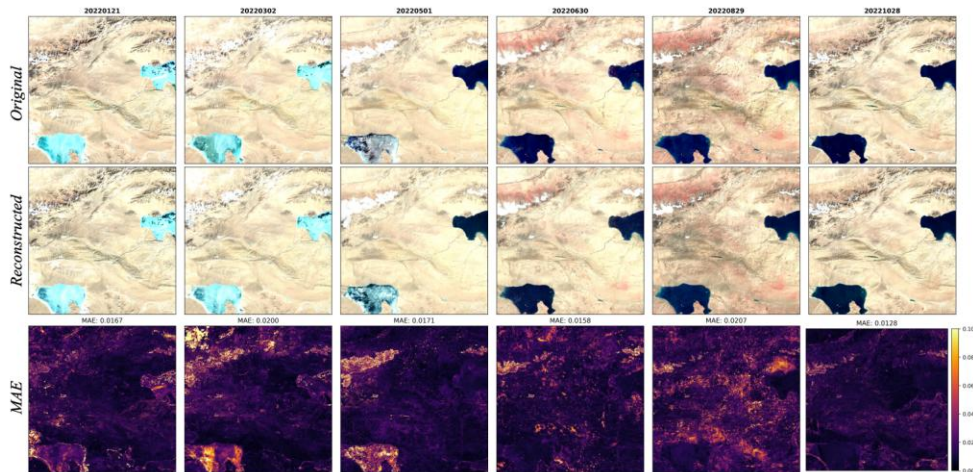


Figure 10: Reconstructive performance in tile 45SXB. The top row displays the original SDC30 surface reflectance (NIR-Red-Green false-color composite). The middle row shows the corresponding reconstructed output from the ESDNet decoder. The bottom row quantifies the spatial distribution of the MAE per pixel.

4.3 Transferability and Generalization Analysis

A primary objective of the Embedded Seamless Data (ESD) product is to provide "universal" representations that support robust generalization beyond the specific supervised tasks encountered during its construction. To evaluate this, we benchmarked the performance of the ESD latent vectors against the original high-dimensional SDC30 reflectance data and standard temporal compositing methods across various downstream classification tasks.

4.3.1 Downstream Classification Performance

We conducted an in-domain transfer experiment using four distinct thematic monitoring tasks: FROMGLC, ESA WorldCover, GLAD Global Crop Extent, and GLAD Global Surface Water. To compare the thematic accuracy of classifications derived from ESD versus raw SDC30 data and Landsat compositing, we employed a suite of common machine learning algorithms, including Linear Prediction Heads, k-Nearest Neighbors (k=1 and k=3), and Random Forests.

As illustrated in Figure 11, using ESD achieves consistently higher performance across all tested algorithms and tasks. Despite a significant reduction in data dimensionality, the ESD latent vectors effectively distill and preserve essential semantic features. A detailed comparison of the confusion matrices for land-cover classification using the Random Forest algorithm is provided in Tables 6 and 7. ESD achieved a higher global OA of 79.74% compared to 76.92% for the raw

SDC30 data. Significant gains were observed in several primary categories. For instance, the Producer’s Accuracy (PA) for Crop improved from 61.82% (SDC30) to 71.96% (ESD), while the PA for Grass rose from 64.01% to 70.25%. The User’s Accuracy (UA) for Impervious surfaces saw a notable increase from 84.38% to 80.62%, and Water reached a high of 95.29%.

These results indicate that the ESD latent space is better organized to resolve spectral-temporal ambiguities, thereby enhancing the classification of complex land cover types.

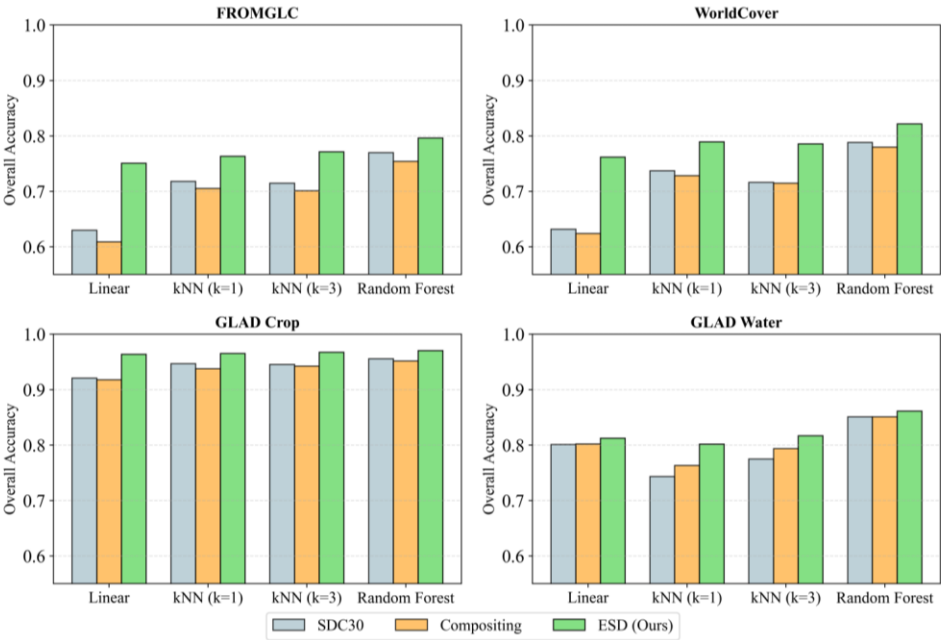


Figure 11: Comparison of downstream classification performance across diverse thematic monitoring tasks. The bar charts illustrate the Overall Accuracy (OA) achieved using the Embedded Seamless Data (ESD) latent representations, the Global 30 m Seamless Data Cube (SDC30) reflectance, and temporal compositing methods. Evaluation is conducted across four distinct tasks: FROMGLC , ESA WorldCover , GLAD Global Crop Extent , and GLAD Global Surface Water. For each task, performance is benchmarked using four classification algorithms: Linear Prediction Head, k-Nearest Neighbors (k=1 and k=3), and Random Forest.

Table 6: Confusion matrix of land-cover classification results using the Global 30 m Seamless Data Cube (SDC30) and Random Forest. The classification was performed on the independent FAST-validation sample set. Producer's Accuracy (PA) and User's Accuracy (UA) are provided for each of the primary land-cover categories.

Class	Crop	Forest	Grass	Shrub	Water	Tundra	Impervious	Bareland	Snow/Ice	PA
Crop	2048	154	808	247	3	0	2	51	0	61.82%
Forest	116	8793	449	320	7	78	0	4	0	90.03%
Grass	362	596	4247	780	33	225	5	387	0	64.01%
Shrub	176	676	907	2669	5	58	3	185	0	57.04%
Water	3	16	24	3	1340	22	0	24	0	93.58%
Tundra	6	69	165	3	18	1196	0	71	0	78.27%
Impervious	57	18	75	25	1	0	54	42	0	19.85%
Bareland	5	7	203	108	24	50	0	5221	7	92.82%
Snow/Ice	0	0	1	0	5	0	0	20	118	81.94%
UA	73.86%	85.13%	61.74%	64.24%	93.31%	73.42%	84.38%	86.94%	94.40%	76.92%

Table 7: Confusion matrix of land-cover classification results using the Embedded Seamless Data (ESD) and Random Forest. The classification was performed on the independent FAST-validation sample set. Producer's Accuracy (PA) and User's Accuracy (UA) are provided for each of the primary land-cover categories.

Class	Crop	Forest	Grass	Shrub	Water	Tundra	Impervious	Bareland	Snow/Ice	PA
Crop	2384	136	609	141	3	0	9	31	0	71.96%
Forest	72	8902	412	319	7	46	5	4	0	91.14%
Grass	322	515	4661	592	25	196	8	316	0	70.25%
Shrub	96	683	923	2766	2	49	1	159	0	59.12%
Water	3	15	35	1	1334	22	1	21	0	93.16%
Tundra	0	65	208	4	12	1159	0	80	0	75.85%
Impervious	24	24	70	19	2	0	104	29	0	38.24%
Bareland	7	7	242	91	12	49	1	5210	6	92.62%
Snow/Ice	0	0	0	0	3	0	0	31	110	76.39%
UA	81.98%	86.03%	65.10%	70.33%	95.29%	76.20%	80.62%	88.59%	94.83%	79.74%

540 4.3.2 Few-Shot Learning Capability

Sample size is important in global land cover mapping (Gong et al., 2024, 2019). To assess the practical utility of ESD in data-scarce environments, we conducted a few-shot learning examination on the FAST-validation set. We compared the performance of ESD and SDC30 across a range of training sample sizes, from 10^2 to nearly 10^5 samples.

545 As illustrated in **Figure 12**, ESD consistently outperforms SDC30 in data-scarce regimes. Specifically, the ESD-based linear classifier reaches a performance plateau much earlier in the sample size progression compared to SDC30, which struggles to maintain accuracy at lower sample counts. Even with only 100 training samples, ESD-based models provide significantly higher Overall Accuracy than their raw-data counterparts. This demonstrates that the ESD latent space is already pre-organized into meaningful semantic clusters. This few-shot advantage is maintained across multiple algorithms, including KNN and Random Forest, highlighting the "universal" nature of the learned representations and their adaptability
550 to emergent downstream applications in Earth system science.

By reducing the requirement for massive labeled datasets, the ESD product enables more agile and cost-effective environmental monitoring, particularly for regions or phenomena where ground-truth data is difficult to obtain

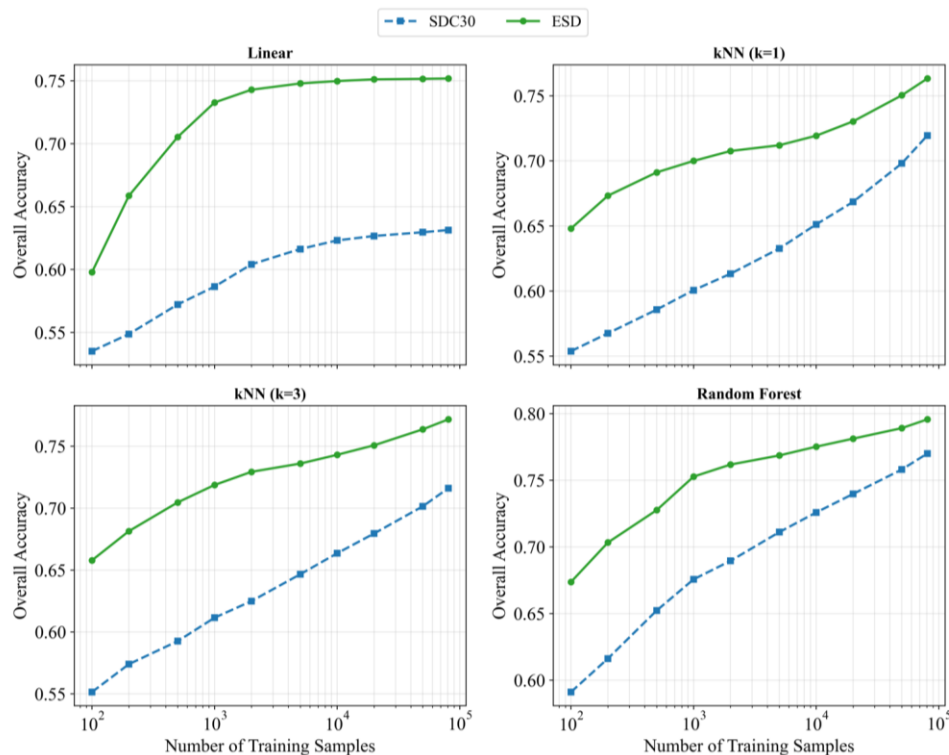


Figure 12: Comparison of downstream classification performance across diverse thematic monitoring tasks. The bar charts illustrate the Overall Accuracy (OA) achieved using the Embedded Seamless Data (ESD) latent representations, the Global 30 m Seamless Data Cube (SDC30) reflectance, and temporal compositing methods. Evaluation is conducted across four distinct tasks: FROMGLC , ESA WorldCover , GLAD Global Crop Extent , and GLAD Global Surface Water. For each task, performance is benchmarked using four classification algorithms: Linear Prediction Head, k-Nearest Neighbors (k=1 and k=3), and Random Forest.

5. Discussion

5.1 Implicit Denoising Effects via the Encoder-Decoder Process

A significant emergent property of the ESDNet architecture is its inherent capacity to "denoise" satellite time-series data during the latent encoding and reconstruction stages. While the primary goal of the framework is dimensionality reduction and feature extraction, the process of bottlenecking high-dimensional observations into a compact, quantized latent space effectively filters out transient non-geophysical noise. By projecting raw spectral data into a unified, lower-dimensional latent space, the model is forced to prioritize the persistent biophysical signals (such as vegetation phenology and underlying surface reflectance) over sporadic atmospheric interference. This phenomenon aligns with established deep learning principles where a constrained bottleneck prevents the network from memorizing stochastic noise, instead compelling it to learn the most salient and stable hierarchical patterns of the terrestrial surface.

The practical utility of this denoising capability is clearly illustrated in the various case studies provided in **Figure 13**. In the first three columns of the figure, raw SDC30 images contain persistent cloud residuals (common artifacts that often survive traditional cloud-masking pipelines), which are effectively mitigated in the images decoded from the ESD latent space. The reconstruction process leverages the temporal depth of the embeddings to "fill in" these corrupted pixels with physically consistent data derived from the annual phenological cycle. Similarly, the final two columns of **Figure 13** highlight instances where cloud shadows in the raw SDC30 data were successfully compressed during the ESDNet decoding process.

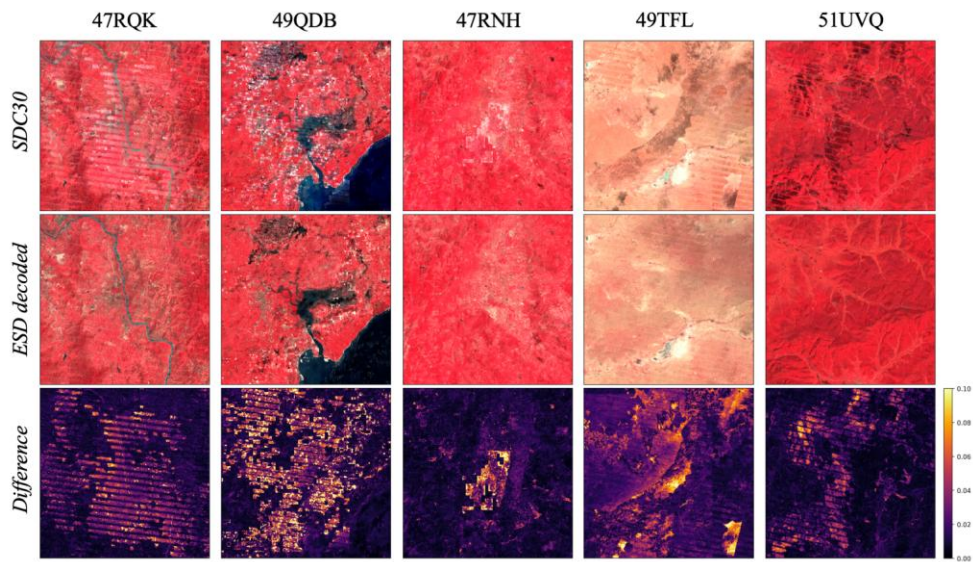


Figure 13: Examples of ESD denoising effects. The first three columns show examples that the cloud residuals in raw SDC30 were effectively mitigated in the images decoded from the ESD. The last two columns highlight instances where cloud shadows in the raw SDC30 data were successfully compressed.

5.2 Ablation study

To determine the optimal configuration for the Embedded Seamless Data (ESD) product, we conducted a systematic ablation study focusing on the core architectural hyperparameters of ESDNet. These experiments evaluate the trade-offs between reconstructive fidelity (measured by MAE, RMSE, and CC) and downstream semantic performance (Overall Accuracy on the ESA WorldCover task using a linear head), while considering the practical constraints of global-scale data storage.

The temporal dimension of the embeddings represents the number of latent steps used to summarize the annual phenological cycle. As shown in **Table 8**, increasing the temporal steps from 4 to 24 results in a monotonic improvement in reconstruction accuracy (MAE decreases from 0.0178 to 0.0101). However, higher temporal dimensionality directly increases the storage footprint of the database. We selected a temporal dimension of 12 as the final configuration. This choice serves two primary functions: a) It achieves high reconstructive fidelity (MAE = 0.0130) and competitive downstream performance (76.4%) while maintaining the ultra-lightweight 2.4 TB/year target; b) It naturally aligns the latent steps with the 12-month seasonal cycle, providing an intuitive structure for researchers analyzing intra-annual land surface dynamics.

Table 8: Performance comparison of varying temporal dimensions for the Embedded Seamless Data (ESD) product. Metrics include reconstruction accuracy (MAE, RMSE, CC) and downstream classification accuracy (OA) for the ESA WorldCover task. The asterisk (*) denotes the parameter (12) selected for the final product to align with the annual monthly cycle.

Temporal dimension	MAE↓	RMSE↓	CC↑	OA↑ (WorldCover, Linear)
4	0.0178	0.0246	0.7652	75.5%
6	0.0162	0.0224	0.7937	75.3%
8	0.0145	0.0200	0.8309	76.2%
12*	0.0130	0.0179	0.8537	76.4%
24	0.0101	0.0138	0.9035	76.8%

The size of the discrete latent space (the virtual codebook) determines the "vocabulary" available to the model to represent Earth's diversity. Experimental results in **Table 9** show that expanding this codebook from 256 to 65,536 significantly improves both spectral reconstruction and land-cover classification accuracy. We finalized the codebook size at 65,536 (2^{16}). This specific value is strategically chosen to allow each embedding value to be stored in a single uint16 format. This hardware-efficient encoding minimizes storage overhead without the semantic degradation observed at lower bit-depths, ensuring the "democratization" of the data for users with limited computational resources.

Table 9: Performance metrics across different Finite Scalar Quantization (FSQ) codebook sizes. The asterisk (*) indicates the codebook size (65,536) used for the final ESD product, chosen to balance representational accuracy with efficient uint16 data storage.

Codebook sizes	MAE↓	RMSE↓	CC↑	OA↑ (WorldCover, Linear)
256	0.0159	0.0215	0.8056	71.4%
1024	0.0145	0.0198	0.8271	74.3%
4096	0.0142	0.0193	0.8364	74.6%
16384	0.0136	0.0187	0.8486	75.1%
65536*	0.0131	0.0180	0.8543	77.3%

The complexity of feature extraction was evaluated by varying the number of residual layers within the encoder and decoder. As indicated in **Table 10**, performance initially improves with depth but begins to plateau and eventually decline after 10 layers. The 10-layer configuration achieved the highest downstream accuracy (77.3%) while keeping reconstruction errors low. Increasing the depth to 20 layers resulted in a slight decrease in accuracy (76.9%), likely due to diminishing returns in feature abstraction or increased optimization complexity. Consequently, 10 residual layers were selected to maximize performance efficiency.

620 **Table 10: Performance metrics across different numbers of residual Conv1D layers. The asterisk (*) highlights the 10-layer configuration chosen for the final product to optimize the trade-off between model accuracy and computational overhead.**

Num of residual layers	MAE↓	RMSE↓	CC↑	OA↑ (WorldCover, Linear)
0	0.0146	0.0198	0.8348	72.0%
1	0.0132	0.0181	0.8538	75.8%
2	0.0126	0.0173	0.8641	75.5%
5	0.0127	0.0174	0.8633	75.2%
10*	0.0131	0.0180	0.8543	77.3%
20	0.0133	0.0183	0.8497	76.9%

625 A pivotal finding of our ablation study is the impact of supervised signals during training. As shown in **Table 11**, an unsupervised model (trained solely on reconstruction) achieves superior reconstruction metrics (MAE = 0.0073) but fails significantly in downstream classification (OA = 60.8%). In contrast, the multi-task supervision framework intentionally introduces a constraint on the latent space. While this slightly increases reconstruction error (MAE = 0.0131), it boosts downstream accuracy to 76.2%. This trade-off confirms that our multi-task strategy is essential for imbuing the ESD embeddings with the semantic richness required for "universal" representation learning, moving beyond simple data compression.

630 **Table 11: Performance metrics with and without multi-task supervisory training.**

Supervisory training	MAE↓	RMSE↓	CC↑	OA↑ (WorldCover, Linear)
Without	0.0073	0.0103	0.9389	60.8%
With*	0.0131	0.0180	0.8543	76.2%

5.3 Comparison with Existing Earth Embedding Databases

635 The landscape of Earth observation has recently transitioned toward pre-computed embedding products to bypass the high inference costs of Geospatial Foundation Models (GFMs) (Brown et al., 2025; Fang et al., 2026). While several pioneering datasets have been released, their utility for longitudinal Earth system research is often constrained by temporal depth or structural inconsistencies. As detailed in **Table 12**, existing products like Major TOM (Francis and Czerkawski, 2024), Clay (Clay, 2024), and Copernicus-Embed (Wang et al., 2025) focus on patch-level descriptors optimized for search and retrieval rather than dense mapping. These datasets typically provide "snapshot" representations from limited time windows, which precludes the analysis of multi-decadal land surface dynamics. Pixel-level alternatives such as the Google Satellite

640 Embedding (Brown et al., 2025) and TESSERA (Feng et al., 2025) offer higher spatial precision but vary in their handling of temporal structures.

Table 12: Technical comparison between ESD and existing Earth embedding databases.

Feature	Major TOM Embeddings	LGND Clay Embeddings	Copernicus-Embed	Tessera Embeddings	Google Satellite Embedding	Embedded Seamless Data (Ours)
Granularity	Patch-level	Patch-level	Patch-level	Pixel-level	Pixel-level	Pixel-level
Spatial Res.	2.14-3.56 km	2.56 km	0.25°	10 m	10 m	30 m
Temporal Span	2017-2024	2024-2025	2021	2024	2017-2024	2000-2024
Temporal Res.	Snapshot	Snapshot	Annual	Annual	Annual	Monthly
Dimensions	2048	1024	768	128	64	12
Dtype/Format	float32	float32	float32	int8	int8	uint16

批注 [JL14]:

645 The Embedded Seamless Data (ESD) distinguishes itself through three primary technical dimensions. First, while most current products utilize 32-bit float formats that demand significant storage, ESD employs Finite Scalar Quantization (FSQ) to store embeddings as hardware-efficient uint16 integers, achieving a ~340-fold reduction in volume. This allows a single year of global 30-m data to be encapsulated in 2.4 TB, whereas raw daily archives would reach petabyte scales. Second, ESD provides the most extensive longitudinal coverage, spanning 25 years (2000–2024), compared to the shorter 5- to 8-year spans of most Sentinel-2 based products. Third, unlike models that collapse temporal information into a single vector, ESD preserves 12 temporal latent steps to explicitly represent intra-annual phenological cycles. However, a limitation of the current ESD product is its 30-m spatial resolution, which, while optimized for decadal studies, is coarser than the 10-m resolution offered by recent Sentinel-2 based embeddings like TESSERA or the Google Satellite Embedding.

5.4 Limitations and Future Directions

655 While the Embedded Seamless Data (ESD) dataset provides a high-fidelity and ultra-lightweight framework for global Earth monitoring, several inherent limitations remain that define the scope of its current application and suggest pathways for future development.

660 **Current Constraints in Thematic Representation:** A primary limitation involves the trade-off between reconstruction fidelity and semantic richness. As demonstrated in our ablation studies, incorporating supervisory signals for multi-task training slightly increases the reconstruction error compared to a purely self-supervised model. This suggests that the latent space must balance "memorizing" raw spectral-temporal signatures with "clustering" according to high-level thematic categories like those found in the WorldCover or FROMGLC datasets. Furthermore, while the current embeddings

effectively capture the 12-month phenological cycle, they may struggle to represent rapid, non-cyclical events (such as sudden flash floods or abrupt wildfire disturbances) that occur at a temporal scale finer than the current latent step resolution.

Potential for Multi-Modal Integration: The current architecture is primarily optimized for optical and topographic data from the Landsat, MODIS, and NASADEM archives. A promising future direction involves the integration of multi-modal observations, particularly Synthetic Aperture Radar (SAR) data from missions like Sentinel-1. Incorporating SAR would enhance the dataset's utility in regions with persistent cloud cover and provide structural information that complements the spectral-temporal features currently captured by the ESDNet encoder, particularly for urban areas and forests.

Scaling Toward Higher Spatial Resolution: Expanding the ESD framework to 10-m resolution by integrating the Sentinel-2 constellation represents another critical frontier. While the current 30-m resolution is sufficient for decadal-scale global studies, 10-m embeddings would significantly improve the characterization of fragmented landscapes, small-scale agriculture, and complex urban morphologies. Such an expansion would require addressing the substantial increase in data volume, potentially through more advanced quantization schemes or hierarchical embedding strategies.

Adaptive and Foundation Model Interoperability: The static nature of the current quantization codebook, while stable, limits the model's ability to adapt to entirely new sensor types without retraining. Future iterations could explore the use of dynamic codebooks or cross-modal foundation models that allow the ESD to serve as a universal "plug-and-play" latent representation for an even broader range of downstream Earth system science tasks. By continuing to refine these information-dense vectors, we aim to further lower the barriers to entry for planetary-scale environmental analysis.

Outlook on Integration with Geospatial Foundation Models: Beyond its standalone use, ESD may also be useful within the emerging geospatial foundation-model ecosystem. Because each pixel-year is represented by 12 quantized temporal tokens stored in uint16 format, ESD provides a compact and temporally structured representation that is substantially lighter than the original reflectance time series while still preserving seasonal information. This makes it a natural candidate for use as a compact temporal input to sequence-based or token-based models, as a target representation in distillation settings, or as a frozen feature space for sparse-label transfer and few-shot downstream applications over long historical periods. The discrete token structure induced by the FSQ bottleneck may also be compatible with masked-token pretraining objectives.

批注 [JL15]:

6. Code and data availability

The Embedded Seamless Data (ESD) product, spanning the global land surface from 2000 to 2024, is publicly accessible via the iEarth platform (Credentials: shuangchen / SDC@2024test) at <https://data-starcloud.pcl.ac.cn/iearthdata/64> and <https://doi.org/10.12436/iEarth.0000.20251229.000064.v1> (Chen, 2025). The dataset is organized by year and adheres to the standard MGRS tiling convention to ensure seamless integration with existing Earth Observation workflows. To facilitate the use of these embeddings, all source code and example scripts for downstream task inference are available at

695 <https://github.com/shuangchence/ESD>. We encourage the scientific community to utilize these resources for efficient,
planetary-scale environmental monitoring and the development of novel geospatial AI applications.

7. Conclusion

700 In this study, we introduced the Embedded Seamless Data (ESD), a first-of-its-kind, ultra-lightweight 30-m Earth embedding
database designed for planetary-scale environmental monitoring. By transforming a 25-year record of global land surface
reflectance into information-dense, quantized latent vectors, the ESD achieves a transformative ~340-fold reduction in data
volume. The entire global land surface for a single year is encapsulated within just 2.4 TB, effectively democratizing
petabyte-scale analysis for researchers without access to massive cloud computing infrastructure.

705 Our evaluation of the ESDNet architecture demonstrates that these embeddings maintain high reconstructive fidelity
across the spectral dimension, with an MAE of 0.0130 and consistent stability from 2000 to 2024. Beyond mere compression,
the multi-task learning framework ensures that the latent space is semantically organized, significantly outperforming raw
reflectance data in downstream tasks such as land-cover classification, achieving an even higher overall accuracy of 79.74%
than the 76.92% obtained using raw sensor fusion data, and demonstrating superior performance in few-shot learning
scenarios.

710 The ESD product also exhibits an emergent denoising capability, successfully mitigating residual cloud artifacts and
shadows to provide a cleaner, more stable proxy for surface processes. By providing a continuous, longitudinal record that
preserves intra-annual phenological dynamics, the ESD offers a robust foundation for tracking decadal shifts in the Earth
system. We believe this dataset represents a significant step toward agile Earth system research, enabling global analyses to
be executed with the same ease as local studies and fostering new frontiers in geospatial AI.

Acknowledgements

715 This research was jointly supported by the Research Grants Council (RGC) of Hong Kong, China (STG2/P-705/24-R,
HKU17616825), the Croucher Foundation (CAS22902/CAS22HU01), Hong Kong, China, and the Major Key Project of
PCL (PCL2024A04).

Author contribution

720 SC conceptualized the research idea, developed the methodologies, and performed the formal analysis. SC and YL
conducted the validation. SC, SY, and JL were responsible for visualization. SY, XX, and JBW curated the data. JW and JY
configured and maintained the computing resources. XX and JBW developed the web-based user interface. JW and PG
provided supervision. PG was responsible for resources, funding acquisition, and project administration. SC prepared the
manuscript with contributions from all co-authors. All authors reviewed and edited the manuscript.

725

Competing interests. At least one of the (co-)authors is a member of the editorial board of Earth System Science Data.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Abowarda, A.S., Bai, L., Zhang, C., Long, D., Li, X., Huang, Q., Sun, Z., 2021. Generating surface soil moisture at 30 m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale. *Remote Sensing of Environment* 255, 112301. <https://doi.org/10.1016/j.rse.2021.112301>
- Aguirre-Gutiérrez, J., Berenguer, E., Oliveras Menor, I., Bauman, D., Corral-Rivas, J.J., Nava-Miranda, M.G., Both, S., Ndong, J.E., Ondo, F.E., Bengone, N.N., Mihinhou, V., Dalling, J.W., Heineman, K., Figueiredo, A., González-M, R., Norden, N., Hurtado-M, A.B., González, D., Salgado-Negret, B., Reis, S.M., Moraes de Seixas, M.M., Farfan-Rios, W., Shenkin, A., Riutta, T., Girardin, C.A.J., Moore, S., Abernethy, K., Asner, G.P., Bentley, L.P., Burslem, D.F.R.P., Cernusak, L.A., Enquist, B.J., Ewers, R.M., Ferreira, J., Jeffery, K.J., Joly, C.A., Marimon-Junior, B.H., Martin, R.E., Morandi, P.S., Phillips, O.L., Bennett, A.C., Lewis, S.L., Quesada, C.A., Marimon, B.S., Kissling, W.D., Silman, M., Teh, Y.A., White, L.J.T., Salinas, N., Coomes, D.A., Barlow, J., Adu-Bredu, S., Malhi, Y., 2022. Functional susceptibility of tropical forests to climate change. *Nat Ecol Evol* 6, 878–889. <https://doi.org/10.1038/s41559-022-01747-6>
- Bengio, Y., Léonard, N., Courville, A., 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. <https://doi.org/10.48550/arXiv.1308.3432>
- Bolton, D.K., Gray, J.M., Melaas, E.K., Moon, M., Eklundh, L., Friedl, M.A., 2020. Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery. *Remote Sensing of Environment* 240, 111685. <https://doi.org/10.1016/j.rse.2020.111685>
- Boyte, S.P., Wylie, B.K., Rigge, M.B., Dahal, D., 2018. Fusing MODIS with Landsat 8 data to downscale weekly normalized difference vegetation index estimates for central Great Basin rangelands, USA. *GIScience & Remote Sensing* 55, 376–399. <https://doi.org/10.1080/15481603.2017.1382065>
- Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., Tait, A.M., 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci Data* 9, 251. <https://doi.org/10.1038/s41597-022-01307-4>
- Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L.L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., 2025. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. <https://doi.org/10.48550/arXiv.2507.22291>
- Chen, B., Chen, L., Huang, B., Michishita, R., Xu, B., 2018. Dynamic monitoring of the Poyang Lake wetland by integrating Landsat and MODIS observations. *ISPRS Journal of Photogrammetry and Remote Sensing* 139, 75–87. <https://doi.org/10.1016/j.isprsjprs.2018.02.021>
- Chen, S., 2025. Embedded Seamless Data: An ultra-lightweight Earth embedding database for accurate and flexible global land monitoring. <https://doi.org/10.12436/iEarth.0000.20251229.000064.v1>
- Chen, S., Wang, J., Gong, P., 2023. ROBOT: A spatiotemporal fusion model toward seamless data cube for global remote sensing applications. *Remote Sensing of Environment* 294, 113616. <https://doi.org/10.1016/j.rse.2023.113616>
- Chen, S., Wang, J., Liu, Q., Liang, X., Liu, R., Qin, P., Yuan, J., Wei, J., Yuan, S., Huang, H., Gong, P., 2024. Global 30 m seamless data cube (2000–2022) of land surface reflectance generated from Landsat 5, 7, 8, and 9 and MODIS Terra constellations. *Earth Syst. Sci. Data* 16, 5449–5475. <https://doi.org/10.5194/essd-16-5449-2024>

- 770 Chuvieco, E., Ventura, G., Martín, M.P., Gómez, I., 2005. Assessment of multitemporal compositing techniques of MODIS and AVHRR images for burned land mapping. *Remote Sensing of Environment* 94, 450–462. <https://doi.org/10.1016/j.rse.2004.11.006>
- Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment* 219, 145–161. <https://doi.org/10.1016/j.rse.2018.09.002>
- 775 Clay, 2024. Clay Foundation Model [WWW Document]. URL <https://github.com/Clay-foundation/model>
- Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S., 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery.
- Czerkawski, M., Kluczek, M., Bojanowski, J.S., 2024. Global and Dense Embeddings of Earth: Major TOM Floating in the Latent Space. <https://doi.org/10.48550/arXiv.2412.05600>
- 780 Declaro, A., Kanae, S., 2024. Enhancing Surface Water Monitoring through Multi-Satellite Data-Fusion of Landsat-8/9, Sentinel-2, and Sentinel-1 SAR. *Remote Sensing* 16, 3329. <https://doi.org/10.3390/rs16173329>
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>
- 785 Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* 120, 25–36. <https://doi.org/10.1016/j.rse.2011.11.026>
- Fang, H., Stewart, A.J., Corley, L., Zhu, X.X., Azizpour, H., 2026. Earth Embeddings as Products: Taxonomy, Ecosystem, and Standardized Access. <https://doi.org/10.48550/arXiv.2601.13134>
- 790 Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M.C., Immitzer, M., Jackson, T., Ball, J., Coomes, D.A., Madhavapeddy, A., Blake, A., Keshav, S., 2025. TESSERA: Temporal Embeddings of Surface Spectra for Earth Representation and Analysis. <https://doi.org/10.48550/arXiv.2506.20380>
- Forzieri, G., Dakos, V., McDowell, N.G., Ramdane, A., Cescatti, A., 2022. Emerging signals of declining forest resilience under climate change. *Nature* 608, 534–539. <https://doi.org/10.1038/s41586-022-04959-9>
- 795 Francis, A., Czerkawski, M., 2024. Major TOM: Expandable Datasets for Earth Observation, in: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Athens, Greece, pp. 2935–2940. <https://doi.org/10.1109/IGARSS53475.2024.10640760>
- 800 Gao, F., Masek, J., Schwaller, M., Hall, F., 2006. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. *IEEE Trans. Geosci. Remote Sensing* 44, 2207–2218. <https://doi.org/10.1109/TGRS.2006.872081>
- Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., Zhou, Y., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment* 236, 111510. <https://doi.org/10.1016/j.rse.2019.111510>
- 805 Gong, P., Liang, S., Carlton, E.J., Jiang, Q., Wu, J., Wang, L., Remais, J.V., 2012. Urbanisation and health in China. *The Lancet* 379, 843–852. [https://doi.org/10.1016/S0140-6736\(11\)61878-3](https://doi.org/10.1016/S0140-6736(11)61878-3)
- Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, Wenyu, Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, Weijia, Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Science Bulletin* 64, 370–373. <https://doi.org/10.1016/j.scib.2019.03.002>
- 810 Gong, P., Wang, J., Huang, H., 2024. Stable classification with limited samples in global land cover mapping: Theory and experiments. *Science Bulletin* 69, 1862–1865. <https://doi.org/10.1016/j.scib.2024.03.040>
- 815 Gong, P., Wang, J., Yu, L., Zhao, Yongchao, Zhao, Yuanyuan, Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, Han, Zhu, P., Zhao, Z., Zhang, Haiying, Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, Liang, Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, Jin, Chen, Jun, 2013. Finer resolution observation

and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing* 34, 2607–2654. <https://doi.org/10.1080/01431161.2012.748992>

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>

Goyena, H., Pérez-Goya, U., Montesino-SanMartin, M., Militino, A.F., Wang, Q., Atkinson, P.M., Ugarte, M.D., 2023. Unpaired spatio-temporal fusion of image patches (USTFIP) from cloud covered images. *Remote Sensing of Environment* 295, 113709. <https://doi.org/10.1016/j.rse.2023.113709>

Guo, D., Shi, W., Hao, M., Zhu, X., 2020. FSDAF 2.0: Improving the performance of retrieving land cover changes and preserving spatial details. *Remote Sensing of Environment* 248, 111973. <https://doi.org/10.1016/j.rse.2020.111973>

Guo, X., Lao, J., Dang, B., Zhang, Yingying, Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Yongjun, Li, Y., 2024. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 27662–27673. <https://doi.org/10.1109/CVPR52733.2024.02613>

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* 342, 850–853. <https://doi.org/10.1126/science.1244693>

Hansen, M.C., Roy, D.P., Lindquist, E., Adusei, B., Justice, C.O., Altstatt, A., 2008. A method for integrating MODIS and Landsat data for systematic monitoring of forest cover and change in the Congo Basin. *Remote Sensing of Environment* 112, 2495–2513. <https://doi.org/10.1016/j.rse.2007.11.012>

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R., 2022. Masked Autoencoders Are Scalable Vision Learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 15979–15988. <https://doi.org/10.1109/CVPR52688.2022.01553>

Hilker, T., Wulder, M.A., Coops, N.C., Linke, J., McDermid, G., Masek, J.G., Gao, F., White, J.C., 2009. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sensing of Environment* 113, 1613–1627. <https://doi.org/10.1016/j.rse.2009.03.007>

Huang, X., Song, Y., Yang, J., Wang, W., Ren, H., Dong, M., Feng, Y., Yin, H., Li, J., 2022. Toward accurate mapping of 30-m time-series global impervious surface area (GISA). *International Journal of Applied Earth Observation and Geoinformation* 109, 102787. <https://doi.org/10.1016/j.jag.2022.102787>

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* 83, 195–213. [https://doi.org/10.1016/S0034-4257\(02\)00096-2](https://doi.org/10.1016/S0034-4257(02)00096-2)

Ienco, D., Interdonato, R., Gaetano, R., Ho Tong Minh, D., 2019. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS Journal of Photogrammetry and Remote Sensing* 158, 11–22. <https://doi.org/10.1016/j.isprsjprs.2019.09.016>

Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S.K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R., 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. <https://doi.org/10.48550/arXiv.2310.18660>

Ji, L., Gong, P., Wang, J., Shi, J., Zhu, Z., 2018. Construction of the 500-m Resolution Daily Global Surface Water Change Database (2001–2016). *Water Resour. Res.* 54. <https://doi.org/10.1029/2018WR023060>

Klemmer, K., Rolf, E., Rußwurm, M., Camps-Valls, G., Ermon, S., Francis, A., Jacobs, N., Kerner, H., Mackey, L., Mai, G., Aodha, O.M., Reichstein, M., Robinson, C., Rolnick, D., Sitzmann, V., Tuia, D., Zhu, X., 2025. Earth Embeddings: Towards AI-centric Representations of our Planet.

Li, C., Gong, P., Wang, J., Zhu, Z., Biging, G.S., Yuan, C., Hu, T., Zhang, H., Wang, Q., Li, X., Liu, X., Xu, Y., Guo, J., Liu, C., Hackman, K.O., Zhang, M., Cheng, Y., Yu, L., Yang, J., Huang, H., Clinton, N., 2017. The first all-season sample set for mapping global land cover with Landsat-8 data. *Science Bulletin* 62, 508–515. <https://doi.org/10.1016/j.scib.2017.03.011>

Liu, H., Gong, P., Wang, J., Wang, X., Ning, G., Xu, B., 2021. Production of global daily seamless data cubes and quantification of global land cover change from 1985 to 2020 - iMap World 1.0. *Remote Sensing of Environment* 258, 112364. <https://doi.org/10.1016/j.rse.2021.112364>

Liu, S., Zhou, J., Qiu, Y., Chen, J., Zhu, X., Chen, H., 2022. The FIRST model: Spatiotemporal fusion incorporating spectral autocorrelation. *Remote Sensing of Environment* 279, 113111. <https://doi.org/10.1016/j.rse.2022.113111>

Manas, O., Lacoste, A., Giro-i-Nieto, X., Vazquez, D., Rodriguez, P., 2021. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 9394–9403. <https://doi.org/10.1109/ICCV48922.2021.00928>

Markham, B.L., Helder, D.L., 2012. Forty-year calibrated record of earth-reflected radiance from Landsat: A review. *Remote Sensing of Environment* 122, 30–40. <https://doi.org/10.1016/j.rse.2011.06.026>

Masek, J.G., Wulder, M.A., Markham, B., McCorkel, J., Crawford, C.J., Storey, J., Jenstrom, D.T., 2020. Landsat 9: Empowering open science and applications through continuity. *Remote Sensing of Environment* 248, 111968. <https://doi.org/10.1016/j.rse.2020.111968>

Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M., 2023. Finite Scalar Quantization: VQ-VAE Made Simple. <https://doi.org/10.48550/arXiv.2309.15505>

Morissette, J.T., Privette, J.L., Justice, C.O., 2002. A framework for the validation of MODIS Land products. *Remote Sensing of Environment* 83, 77–96. [https://doi.org/10.1016/S0034-4257\(02\)00088-3](https://doi.org/10.1016/S0034-4257(02)00088-3)

NASA JPL, 2020. NASADEM Merged DEM Global 1 arc second V001. https://doi.org/10.5067/MEASURES/NASADEM/NASADEM_HGT.001

Oord, A. van den, Vinyals, O., Kavukcuoglu, K., 2018. Neural Discrete Representation Learning. <https://doi.org/10.48550/arXiv.1711.00937>

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 418–422. <https://doi.org/10.1038/nature20584>

Piao, S., Liu, Q., Chen, A., Janssens, I.A., Fu, Y., Dai, J., Liu, L., Lian, X., Shen, M., Zhu, X., 2019. Plant phenology and global climate change: Current progresses and challenges. *Glob Change Biol* 25, 1922–1940. <https://doi.org/10.1111/gcb.14619>

Pickens, A.H., Hansen, M.C., Hancher, M., Stehman, S.V., Tyukavina, A., Potapov, P., Marroquin, B., Sherani, Z., 2020. Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. *Remote Sensing of Environment* 243, 111792. <https://doi.org/10.1016/j.rse.2020.111792>

Pickens, A.H., Hansen, M.C., Stehman, S.V., Tyukavina, A., Potapov, P., Zalles, V., Higgins, J., 2022. Global seasonal dynamics of inland open water and ice. *Remote Sensing of Environment* 272, 112963. <https://doi.org/10.1016/j.rse.2022.112963>

Potapov, P., Turubanova, S., Hansen, M.C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., Cortez, J., 2021. Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nat Food* 3, 19–28. <https://doi.org/10.1038/s43016-021-00429-z>

Shang, R., Zhu, Z., 2019. Harmonizing Landsat 8 and Sentinel-2: A time-series-based reflectance adjustment approach. *Remote Sensing of Environment* 235, 111439. <https://doi.org/10.1016/j.rse.2019.111439>

Shang, R., Zhu, Z., Zhang, J., Qiu, S., Yang, Z., Li, T., Yang, X., 2022. Near-real-time monitoring of land disturbance with harmonized Landsats 7–8 and Sentinel-2 data. *Remote Sensing of Environment* 278, 113073. <https://doi.org/10.1016/j.rse.2022.113073>

Smits, M. (Ed.), 2022. Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28 – March 4, 2022, Proceedings, Part I, Lecture Notes in Computer Science. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-030-96957-8>

Soto Vega, P.J., Costa, G.A.O.P. da, Feitosa, R.Q., Ortega Adame, M.X., Almeida, C.A. de, Heipke, C., Rottensteiner, F., 2021. An unsupervised domain adaptation approach for change detection and its application to deforestation

mapping in tropical biomes. *ISPRS Journal of Photogrammetry and Remote Sensing* 181, 113–128. <https://doi.org/10.1016/j.isprsjprs.2021.08.026>

Szwarcman, D., Roy, S., Fraccaro, P., Gislason, P.E., Blumenstiel, B., Ghosal, R., Oliveira, P.H. de, Almeida, J.L. de S., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Gomes, C., Kumar, A., Truong, M., Godwin, D., Lee, H., Hsu, C.-Y., Asanjan, A.A., Mujeci, B., Shidham, D., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Hain, C., Kennedy, R., Zadrozny, B., Bell, D., Cavallaro, G., Watson, C., Maskey, M., Ramachandran, R., Moreno, J.B., 2025. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. <https://doi.org/10.48550/arXiv.2412.02732>

Takida, Y., Shibuya, T., Liao, W., Lai, C.-H., Ohmura, J., Uesaka, T., Murata, N., Takahashi, S., Kumakura, T., Mitsufuji, Y., 2022. SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization. <https://doi.org/10.48550/arXiv.2205.07547>

Tuia, D., Schindler, K., Demir, B., Zhu, X.X., Kochupillai, M., Džeroski, S., Van Rijn, J.N., Hoos, H.H., Del Frate, F., Datcu, M., Markl, V., Le Saux, B., Schneider, R., Camps-Valls, G., 2025. Artificial Intelligence to Advance Earth Observation: A review of models, recent trends, and pathways forward. *IEEE Geosci. Remote Sens. Mag.* 2–25. <https://doi.org/10.1109/MGRS.2024.3425961>

Wang, Y., Xiong, Z., Liu, C., Stewart, A.J., Dujardin, T., Bountos, N.I., Zavras, A., Gerken, F., Papoutsis, I., Leal-Taixé, L., Zhu, X.X., 2025. Towards a Unified Copernicus Foundation Model for Earth Vision. <https://doi.org/10.48550/arXiv.2503.11849>

Wieland, M., Martinis, S., Kiefl, R., Gstaiger, V., 2023. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. *Remote Sensing of Environment* 287, 113452. <https://doi.org/10.1016/j.rse.2023.113452>

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free access to Landsat imagery. *Science* 320, 1011. <https://doi.org/10.1126/%20science.320.5879.1011a>

Wulder, M.A., Roy, D.P., Radeloff, V.C., Loveland, T.R., Anderson, M.C., Johnson, D.M., Healey, S., Zhu, Z., Scambos, T.A., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C.J., Masek, J.G., Hermosilla, T., White, J.C., Belward, A.S., Schaaf, C., Woodcock, C.E., Huntington, J.L., Lymburner, L., Hostert, P., Gao, F., Lyapustin, A., Pekel, J.-F., Strobl, P., Cook, B.D., 2022. Fifty years of Landsat science and impacts. *Remote Sensing of Environment* 280, 113195. <https://doi.org/10.1016/j.rse.2022.113195>

Xu, J., Liang, S., Ma, H., He, T., 2022. Generating 5 km resolution 1981–2018 daily global land surface longwave radiation products from AVHRR shortwave and longwave observations using densely connected convolutional neural networks. *Remote Sensing of Environment* 280, 113223. <https://doi.org/10.1016/j.rse.2022.113223>

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., Arino, O., 2022. ESA WorldCover 10 m 2021 v200.

Zhao, Q., Yu, L., Li, X., Peng, D., Zhang, Y., Gong, P., 2021. Progress and Trends in the Application of Google Earth and Google Earth Engine. *Remote Sensing* 13, 3778. <https://doi.org/10.3390/rs13183778>

Zhu, X., Cai, F., Tian, J., Williams, T., 2018. Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. *Remote Sensing* 10, 527. <https://doi.org/10.3390/rs10040527>

Zhu, X., Chen, J., Gao, F., Chen, X., Masek, J.G., 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment* 114, 2610–2623. <https://doi.org/10.1016/j.rse.2010.05.032>

Zhu, X., Helmer, E.H., Gao, F., Liu, D., Chen, J., Lefsky, M.A., 2016. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sensing of Environment* 172, 165–177. <https://doi.org/10.1016/j.rse.2015.11.016>