Towards Streaming Continual Learning for Earth Observation Multimodal Foundation Models

Marcello M. Declich

[1]*Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Via Leonardo da Vinci, 32, Milan, 20131, Italy*

Abstract

Earth Observation (EO) systems generate continuous multimodal data streams at unprecedented scales. However, in this context, the literature offers solutions based on foundation models that operate within static training paradigms, which limit their effectiveness. Trained once on historical datasets and deployed without further learning, these models face critical issues when confronted with the dynamic nature of the environment, which includes emerging phenomena, sensor degradation, and evolving environmental patterns. This vision paper identifies three fundamental gaps: (1) the absence of memory-efficient anti-forgetting mechanisms at the foundation scale, (2) static cross-modal fusion strategies that cannot adapt to changing observational contexts, and (3) temporal representations that fail to distinguish cyclical patterns from distributional drift. Addressing these limitations requires convergence of foundation models, Continual Learning, and Streaming Machine Learning. This work envisions three key research directions: efficient model updating through selective replay and parameter regularization, explicit drift detection mechanisms, and context-dependent fusion strategies. These directions aim to enable EO systems that continuously learn from terabyte-per-day satellite streams while maintaining transfer learning capabilities and computational feasibility essential for operational deployment.

Keywords

Geospatial Foundation Models, Streaming Continual Learning, Multimodal Fusion, Concept Drift Detection, Earth Observation

1. Introduction

In recent years, the availability of digitized information about the Earth has increased exponentially, opening new opportunities for applying large Artificial Intelligence (AI) analysis models in fields such as environmental monitoring, disaster management and response, urban planning and smart cities, security, and defense applications.

These AI models are inherently *multimodal systems*, meaning that they must process and integrate data from different types of representations. **Earth Observation (EO)** data is inherently multimodal, ranging from high-resolution optical and Synthetic Aperture Radar (SAR) satellite imagery to ground-based sensor networks, from textual reports to elevation and 3D products (e.g., Digital Elevation Models, LiDAR point clouds). In addition, EO include radiometric measurements, i.e., the intensity recorded by sensors across different wavelengths of the electromagnetic spectrum. This diversity particularly benefits from AI approaches that can exploit a complex ecosystem of interconnected information. Data is complementary, and their interconnection enables a richer understanding of the Earth system.

The core challenge of multimodal learning is to combine information across modalities, which is usually referred to as *cross-modal fusion*. For example, data coming from optical sensors is obscured by clouds, but SAR images penetrate them and provide an image of the Earth regardless of weather conditions. At the same time, this diversity amplifies the challenges of generalizing EO data.

This heterogeneity of data is accompanied by diversity in data generation and reception. Data are produced constantly as a continuous datastream (unbounded sequence of data generated and transmitted over time), with each satellite generating terabytes of data daily. The revisit frequency (the time interval between successive observations of the same location by a satellite or satellite constellation) for a single

1st Streaming Continual Learning Bridge at AAAI26, January 21, 2026, Singapore.

marcellomatteo.declich@polimi.it (M. M. Declich)

https://github.com/MarcelloMatteoDeclich (M. M. Declich)

0009-0008-2985-5906 (M. M. Declich)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-03-13

CEUR-WS.org/Vol-4183/paper4.pdf

region varies, ranging from days for polar-orbiting satellites to continuous observation for geostationary satellites. Resolution also varies significantly with spectral bands of Sentinel Satellite capturing data at 10, 20,or 60 m resolution[1]

Moreover, the assumption that data is independent and identically distributed (i.i.d.), which is usually made in traditional machine learning, does not hold in the EO scenario. According to this assumption, each observation must be drawn from the same probability distribution and not influence the others.

A first class of violations of the i.i.d assumption in EO data arises from spatial and temporal dependence among observations. EO measurements are autocorrelated at multiple scales (e.g., diurnal, seasonal, or orbital) and dependent on one another in several complementary ways.

First, dependence may arise across different spatial tiles observed at the same time. Images acquired at nearby locations are often correlated due to shared environmental conditions and similar land cover types (e.g. fields, buildings, trees, etc.), leading to *spatial dependence*.

Second, *temporal dependence* may arise when the same spatial tile is observed at different time points. In this case, repeated acquisitions of the same location may be inherently correlated, as they reflect the temporal evolution of the same underlying scene. Such dependence may be driven by periodic patterns, including diurnal or seasonal cycles, such as vegetation cycles as illustrated in the top rows of Figure 1. However, temporal dependence is not limited to strictly periodic phenomena. Observations collected at adjacent time points tend to be strictly correlated due to gradual and continuous changes in environmental conditions.

Additionally, the observation modalities themselves are non-stationary in their information acquisition characteristics. Dependence may also arise from the sensing process and temporal acquisition patterns, as different modalities can be partially or entirely unavailable depending on the time of day or atmospheric conditions. For instance, optical sensors cannot acquire data during nighttime hours. Furthermore, the correlation structure between modalities is dynamic rather than fixed: observations from different sensors may exhibit strong correlation under certain conditions, such as during clear, sunny weather, while becoming uncorrelated or complementary under others, such as during cloudy conditions when synthetic aperture radar provides information unavailable to optical instruments.

A distinct, yet related, violation of the i.i.d. assumption occurs when the data distribution evolves over time. Such changes are commonly referred to as *concept drift* and may be driven by long-term climate trends, sensor degradation, anthropogenic activities, or extreme events. In this case, the relationship between inputs and their associated outputs is no longer stationary is no longer stationary, as illustrated in the bottom rows of Figure 1.

Drifts are extensively addressed by two research areas: *Continual Learning (CL)* [1] and *Streaming Machine Learning (SML)* [2] from two different perspectives. CL aims to avoid *catastrophic forgetting*, the issue that arises when a model adapts to a new distribution and may lose knowledge it acquired from older distributions. The CL goal is to strike a balance between acquiring new knowledge (*plasticity*) and retaining the past (*stability*). On the other hand, SML specifically focuses on real-time adaptation, often emphasizing the current distribution and potentially ignoring the past. SML also proposes explicit concept drift detectors to detect changes and prevent models from experiencing strong performance decreases. Recently, **Streaming Continual Learning** (SCL) [3, 4, 5] has been introduced as a distinct and unified paradigm that bridges CL and SML to produce complete and integrated solutions with the goals of rapid adaption to drifts and knowledge retention.

Nevertheless, CL and SML approaches have been little studied in EO, and further investigation is needed. The literature mainly reports applications to unimodal cases or simple vision tasks [6], and their extension to complex multimodal terabyte-scale fusion data scenarios needs to be explored.

In recent years, research in EO has focused on the development of **Foundation Models (FM)**. FMs are trained on huge amounts of unlabeled multimodal data using a self-supervised approach. The models do not learn to solve a specific task but instead learn a representation of the data by combining the various data modalities (and thus learn to balance them).

[1]For ESA Sentinel-2 instrument specifications, see https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Instrument

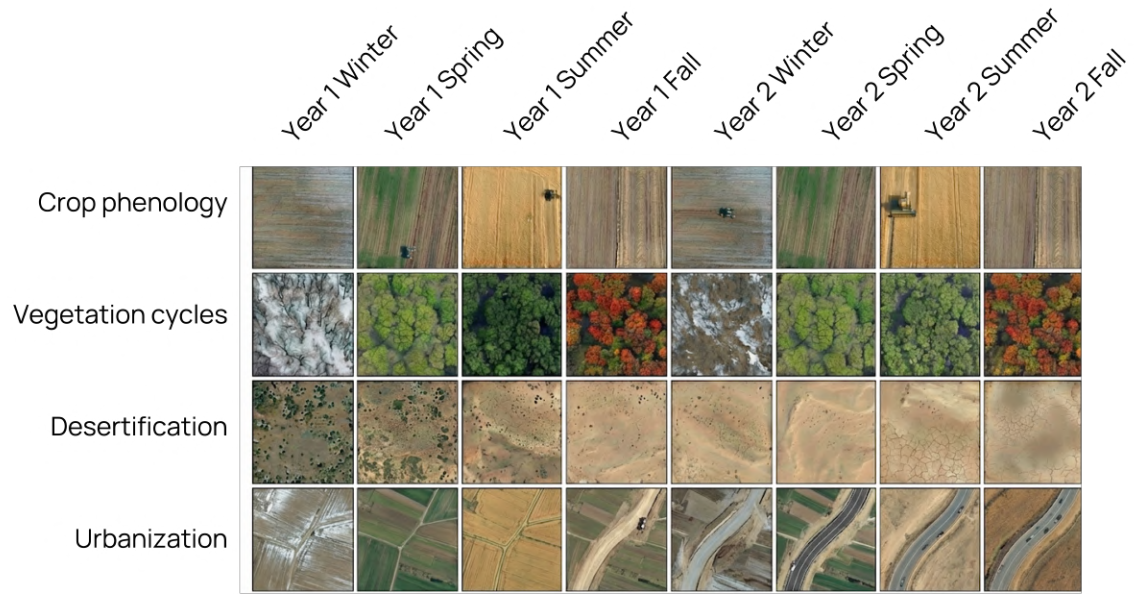


Figure 1: Temporal evolution in Earth Observation imagery. The top two rows illustrate *seasonal cyclic patterns*, where the relationship between inputs and outputs exhibits *temporal dependence* driven by periodic dynamics (e.g., crop phenology and forest vegetation cycles), while the underlying land cover type remains stable. The bottom rows depict *structural changes*, such as desertification trends and infrastructure development (e.g., road construction), where temporal dependence leads to a non-stationary input–output relationship due to permanent transformations of the observed area.

FMs have high transfer learning capabilities, cross-modality understanding of the data, and simplify the implementation of downstream tasks by reducing the amount of task-specific training data, allowing for few-shot adaptation capabilities. However, they operate in a static paradigm that trains the models once on historical data and then deploys them without further learning.

FMs, SML, and CL developed solutions for their own domains, but when applied to operational EO, they are insufficient on their own for creating a system that is simultaneous, multimodal, and drift-aware. Currently, there is no framework in the literature that simultaneously addresses continuous multimodal learning, anti-forgetting applied to FMs, adaptive cross-modal fusion under distribution shift, and explicit detection of concept drift for Earth observation systems. This work calls for the convergence of FM, CL, and SML for operational EO Systems.

This work identifies three critical gaps that prevent existing approaches from scaling to continuous multimodal learning: (1) the absence of memory-efficient anti-forgetting mechanisms at foundation scale, (2) static cross-modal fusion that cannot adapt to changing sensor reliability, (3) temporal representations that conflate cyclical patterns with distributional drift.

The rest of this paper is organized as follows. Section 2 discusses the principles and challenges of CL and SML in detail, while Section 3 provides an overview of geospatial foundation models (GFMs) . Section 4 presents the limitation of the current GFMs while Section 5 presents the proposed research directions. Section 6 presents conclusions and outlines future work.

2. Learning in non-stationary environments

In EO scenarios, data is typically generated continuously at a high frequency. This situation produces what the literature defines as *data stream*, which is formally an unbounded sequence of data points $d_1, d_2, \ldots, d_t, d_{t+1}, \ldots$. When focusing on classification problem, each data point at time t is a couple (X_t, y_t), X_t is a feature vector and y_t is the target label. The assumption is that the real label y_t, associated with d_t, will be available only after receiving X_t and predicting $\hat{y}_t$.

Additionally, the evolving nature of EO data presents non-stationarities. A *concept* is the hidden and

unobservable process that produces the data streem. It is modeled as a stochastic process that generates data points according to the joint distribution $P(X, y)$. A *concept drift* occurs when this probability distribution changes. Concept drift are considered relevant when they require an update of the trained model [7].

The literature distinguishes between two main types of changes [8]. The first, known as **virtual drift**, refers to a change in the input distribution $P(X)$ that affects the feature distribution but not the decision boundary. In EO, for example, a classification task involves a problem such as the seasonal transition from winter to summer, which modifies spectral features but does not alter land cover categories or how previously observed images are classified. A forest remains a forest, but new types of forests with different reflectances may appear throughout the year. While the latter, **real drift**, refers to the case in which the decision boundary (formally defined as the probability $P(y|X)$) itself changes, resulting in an explicit modification of the mapping between inputs and outputs. Following the previous example, recent land-cover standards require a minimum canopy cover to classify an area as forest. Regions that were previously labeled forest may now be labeled shrubland. The appearance may stay similar, but the class meaning changes.

CL assumes a general problem to be split into subproblems called tasks (which correspond to concepts). At each timestamp, the data stream generates a large batch of data (called experience) containing all the data points associated with a distribution. Moving back to the data stream definition, a CL stream is the result of accumulating all data points associated with a specific input distribution $P(X)$ in a batch and presenting them to the model at once. Whenever a new experience is available, the solution has as much time as needed to process it. The goal is to learn how to solve the new task without losing previously acquired knowledge, thus avoiding catastrophic forgetting. Avoiding forgetting in this setting is crucial, as changes introduce a new input distribution without contradicting what has been learned so far.

CL proposes three strategies to avoid catastrophic forgetting [1, 9]: replay-based, regularization, and architectural solutions. Replay-based [10] solutions involve storing a portion of the data observed in previous tasks to be combined with current data during the training phase. Regularization methods [11] work by limiting the loss function during training to preserve performance in previous tasks. Architectural solutions mitigate catastrophic forgetting by adjusting the model's structure as learning progresses. Architectural solutions include freezing weights or neurons, and expanding the network. CL mainly uses Deep Learning approaches.

Conversely, SML does not make any assumptions about the nature of concept drifts, which can be both real and virtual. It usually applies statistical machine learning models (decision trees or probabilistic models) with the goal of focusing just on the current distribution [2]. It is typically applied to tabular data or small computer vision examples. The issue of forgetting is usually overlooked, and the primary objective is to produce models that can quickly adapt to new distributions. Rapid adaptability is, thus, preferred over stability [12]. The main rationale is that if the solution takes a few times to adapt to changes, it does not need to retain past knowledge. Additionally, SML literature proposes concept drift detectors, such as ADWIN [2] and Page-Hinkley [13], methods for detecting when concept drift occurs, and determining when to update the model.

3. Multimodal Geospatial Foundation Model

Multimodal FMs are large-scale AI models, pretrained on vast and heterogeneous data, that learn general-purpose representations and enables efficient adaptation to many downstream tasks across different domains and modalities [14] with minimal additional fine-tuning. FMs provide a generalizable representation framework for enabling cross-task and cross-modal generalization. In the geospatial domain, foundation models inherit this same philosophy but are specifically designed to address the unique characteristics and challenges of remote sensing data. Remote sensing imagery presents fundamental challenges for conventional vision models due to its diverse spatial resolutions, extensive spectral dimensions beyond visible light, and incorporation of specialized data types including radar and LiDAR [14]. The field's requirements for temporal analysis, multi-scale object detection, and

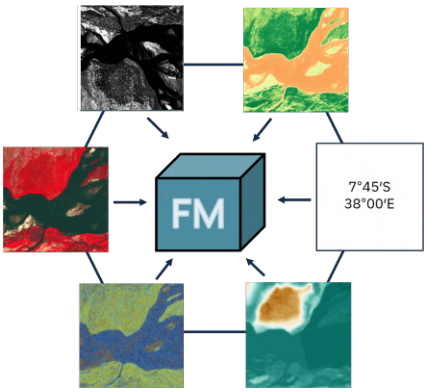


Figure 2: Multimodal Integration in GFMs. The architecture ingests heterogeneous data sources simultaneously to build a unified Earth representation. Inputs include *Synthetic Aperture Radar (SAR)* (top-left), *infrared imagery (IR)* (middle-left), *Normalized Difference Vegetation Index (NDVI)* satellite index that measures the health and density of vegetation based on light reflection (top-right), *Radiometrically Terrain-Correct (RTC)* (bottom-left), *Digital Elevation Models (DEM)* (bottom-right), and *geospatial metadata* such as coordinates (right). Figure adapted from Jakubik et al. [21].

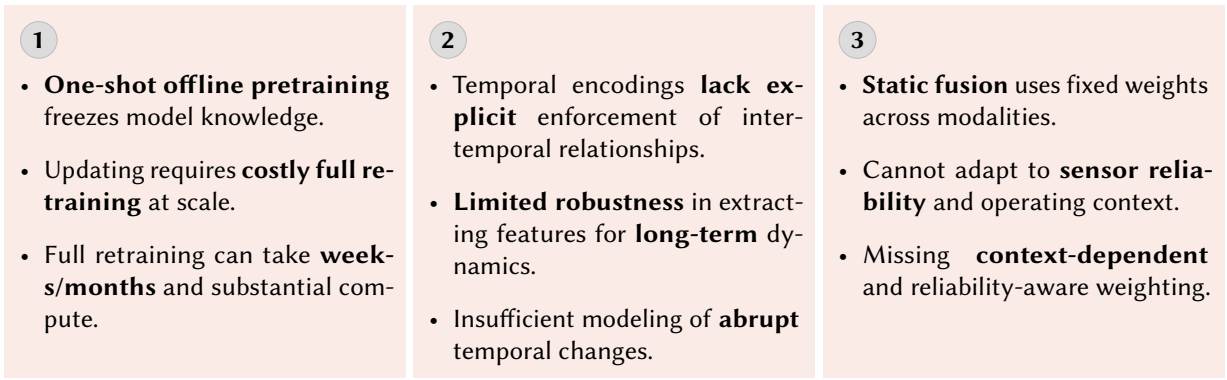
domain-specific tasks create a significant gap between natural image models and geospatial applications. This need has guided scientific research the realization of FMs tailored to remote sensing characteristics.

The evolution of GFMs can be divided into three main generations. The first generation consists of unimodal models based on RGB imagery, trained with self-supervised approaches such as masked image modeling to learn effective representations. The second generation introduces multimodal models that combine data from different sensors, including optical, SAR, multispectral, or LiDAR sources, improving robustness and accuracy. The third and most recent generation is represented by vision-language models that align remote sensing imagery with natural language. This enables advanced applications such as image captioning, visual question answering, and text-based image retrieval [14].

In this paper, we focus on Multimodal GFMs, depicted in Figure 2, however, the underlying reasoning can be extended to Vision-Language GFMs. Multimodal GFMs rely on Transformers [15], such as ViT [16], that are the current standard thanks to their ability to model long-range dependencies, which are essential for large satellite images.

Multimodal GFMs rely on three primary self-supervised training paradigms [14]. Masked Modeling learns general representations by reconstructing randomly masked portions of the input. It is particularly effective for capturing dense spatial contexts and fine-grained details in distinct modalities, as demonstrated by models like SatMAE [17] and RingMo [18]. Contrastive Learning builds discriminative representations by maximizing the similarity between positive pairs (e.g., co-registered optical and SAR patches) while minimizing it for negative ones. This paradigm is the cornerstone for cross-modal alignment and connects visual features with linguistic semantics in Vision-Language models like RemoteCLIP [19]. Finally, Generative Learning models the joint data distribution to generate new samples or reconstruct missing modalities, proving essential for synthesizing complex spatial structures and handling multi-scale hierarchies, such as MetaEarth [20].

Despite the technical complexities, GFMs are widely adopted today because they offer a paradigm shift from isolated, sensor-specific analysis to an integrated, semantic understanding of the Earth [14]. Their key value lies in complementarity: for instance, fusing Optical RGB with SAR allows for all-weather monitoring (seeing through clouds), while integrating LiDAR adds precise geometric structural information unavailable to 2D sensors [14]. Furthermore, these models demonstrate strong zero-shot and few-shot capabilities, enabling adaptation to new tasks with minimal labeled data. In terms of applications, Multimodal GFMs are deployed across diverse downstream tasks categorized into visual and vision-language domains. Visual tasks include Object Detection (e.g., ships, vehicles), Land Cover Classification, and Change Detection for disaster response (e.g., flood assessment) [14].



(a) One-shot pretraining 1 (b) Temporal Encodings 2 (c) Static fusion 3

Figure 3: Key limitations of static geospatial foundation models in non-stationary Earth observation streams.

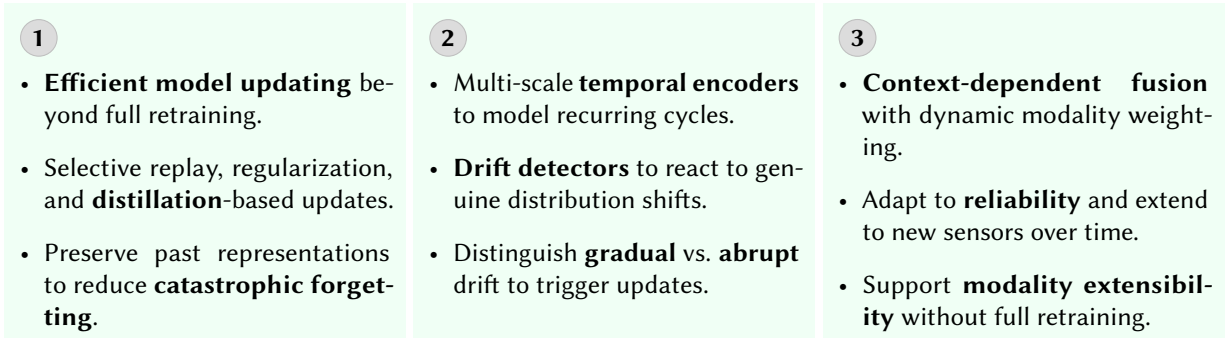
4. Limitations of the Static Paradigm in GFMs

While GFMs have demonstrated remarkable performance in enabling EO downstream tasks (e.g., strong benchmark results on flood detection, burn scar mapping, and crop monitoring), their static training paradigm presents fundamental limitations. These models are typically trained once on historical data, effectively freezing their knowledge at a specific temporal snapshot. This offline approach creates critical bottlenecks for effective operation in dynamic contexts. Three limitations, which are depicted Figure 3, exemplify these constraints and motivate the need for a paradigm shift toward a continuous learning architecture:

1. **One-Shot Offline Pretraining**. GFMs are typically trained only once on huge historical datasets, freezing their knowledge at that specific moment. When new phenomena emerge (e.g., new patterns of deforestation, unprecedented extreme weather events, or sudden changes in land use), the model cannot incorporate this information on its own. Retraining is expensive: updating billions of parameters requires enormous computational resources [22] (e.g., thousands of GPU-hours), prohibitive costs, and weeks or months of time.

2. **Limitations of Temporal Encodings**. Existing GMFs are mainly based [14] on Transformers [15] and they rely on temporal encodings, which can be seen as an evolution of positional encodings [15]. Temporal encodings are learned or predefined vector representations that inject timestamp information into model inputs, enabling the model to associate each observation with its corresponding time of acquisition and to learn how geospatial features change or evolve over time. This design has proven effective, allowing models to develop a meaningful understanding of temporal progression and to jointly reason over spatial information. However, temporal encodings do not explicitly enforce relationships between different time instants within a time series, which may limit the model's ability to robustly extract features that are specifically informative for long-term dynamics and abrupt changes.

3. **Static Cross-Modal Relationships**. The models integrate data from different sensors (optical, radar, thermal, lidar) using attention weights learned during training that remain unchanged. The same fusion strategy is applied regardless of the operating context. In this case, a system that dynamically evaluates the relevance and reliability of each sensor or modality would allow the importance attributed to each source to be remodulated in real time. For example, giving more importance to SAR images in areas with high cloud cover, reducing the importance of a malfunctioning sensor, or excluding certain sensors in extreme weather conditions (e.g., storms, hurricanes, volcanic eruptions). What is missing in these models is methods for learning context-dependent fusion strategies that adapt based on historical performance, current observational conditions, and drift signals.



(a) Efficient updating (b) Temporal structure (c) Adaptive fusion

Figure 4: Research directions towards streaming continual learning for Earth observation foundation models.

5. A Vision for Lifelong GFM Learning

The critical limitations described in the previous section reveal limitations that require a novel research approach at the intersection of FMs, CL, SML, and EO, which simultaneously considers scale, multimodality, non-stationarity, and operational constraints. We envision a new unified framework whose research directions are presented below, with the goal of guiding the community towards systems that can learn continuously from terabyte-per-day multimodal satellite streams while maintaining the transfer learning capabilities and scale that make foundation models powerful. To address these challenges, as illustrated in Figure 4, three interconnected research directions are proposed to enable this vision:

1. **Leverage Efficient Multimodal Foundation Model Updating.** The computational and storage cost of continually updating billions of parameters requires design effective strategies for continuously updating pretrained models, avoiding full model retraining for large GFMs. GFMs should support CL paradigms that allow them to incorporate new information over time while remaining consistent with previously acquired knowledge. This requires mechanisms to mitigate catastrophic forgetting and preserve historical representations. Continual Pretraining [23] is an area of research whith this specific goal. In this direction, CaSSLE [24] and PFR [25] utilize distillation mechanisms to update representations and ensure they are consistent, thereby minimizing the risk of catastrophic forgetting and ensuring they can represent previously observed instances in a meaningful way. It is definitely necessary to research new approaches or adapt existing ones from CL and SML for efficient replay on a multimodal scale, efficient parameter regularization to prevent catastrophic forgetting, and architectural designs tailored to address this specific problem.

2. **Learned Temporal Pattern**. Advancing FMs for EO requires a fundamental shift from passive temporal modeling to explicit temporal structure that can distinguish seasonal patterns from genuine distribution shifts. This evolution must integrate three complementary mechanisms that together create temporally aware embeddings capable of supporting advanced predictive tasks such as crop monitoring, change detection, and disaster response.

   The first mechanism addresses cyclical pattern recognition across multiple temporal scales, including diurnal, weekly, seasonal, and annual rhythms that manifest differently across modalities. Attention-based temporal encoders [26] can learn to recognize these recurring patterns and retrieve appropriate learned representations when familiar cycles reappear, rather than treating each recurrence as a novel task requiring adaptation.

   The second mechanism focuses on utilizing explicit drift detection modules to recognize concept drifts, integrating standard SML drift detectors, such as ADWIN [27] or Page-Hinkley [13].

These detectors can be updated for multimodal EO data to distinguish gradual drift from abrupt shifts. These memory-efficient algorithms prove particularly valuable during disaster scenarios or exceptional events like hurricanes, where rapid response depends on recognizing early indicators of distribution change. Unlike cyclical patterns that require recognition but not continuous updating, trends demand ongoing adaptation mechanisms that can monitor deviations from both historical cycles and expected trajectories. Integrating these adaptation mechanisms directly into pre-trained models would enable continuously updated representations that evolve alongside the temporal dynamics of the observed phenomena.

The third and most transformative mechanism involves moving beyond implicit temporal reasoning by incorporating explicit spatiotemporal masking strategies within FM architectures. By explicitly enforcing interactions between observations across different time istants (e.g, masking image patches across time) models could learn representations inherently sensitive to seasonal dynamics, long-term environmental evolution and abrupt drifts. In this vision, temporal masking is not merely a training heuristic, but a core design principle for GFMs, that associated with a drift detection mechanism, could truly understand Earth system dynamics over time.

3. **Continuous and Context-dependent Cross-Modal Fusion**. During pre-training, FMs typically learn fixed fusion weights to information originating from different modalities. However, a more efficient solution would involve assigning weights dynamically, adapting to the model's varying deployment conditions.

   Modality fusion is critical when applying CL to multi modal sources. During the continuous learning process, features from different modalities (e.g., image and text) tend to diverge and produce a phenomenon referred to as "spatial disorder" [28]. Spatial disorder is the progressive spatial divergence of representations across different modalities that where originally aligned in a shared embedding space and that starts drifting apart. This misalignment leads to more severe performance degradation compared to unimodal models. Consequently, classical multimodal fusion techniques effective in static contexts fail in continuous settings, as different fusion strategies exhibit varying degrees of susceptibility to catastrophic forgetting [28].

   Contemporary GFMs have demonstrated robust performance in integrating multimodal data. However, their reliance on static, pre-defined fusion mechanisms for handling data streams introduces significant limitations. Specifically, current GFMs lack: (1) Adaptative Modality Fusion: a mechanism to dynamically adjust weights to compensate for varying quality and inherent heterogeneity. Approaches such as the *modality fusion network* presented in [29] offer a viable path. Rather than relying on static fusion, a network like this, applied to EO scenarios would allow the model to continuously and adaptively adjust the contribution of each specific modality (e.g.; mitigating the fluctuating utility of optical data under cloud cover). (2) Modality Extensibility: the architectural flexibility to seamlessly incorporate new sensor types or data modalities. Strategies employing meta-learners [30] can facilitate this expansion without incurring catastrophic forgetting or requiring extensive retraining of the entire foundation model.

6. Conclusion

This paper has presented a vision for SCL in EO FMs, addressing fundamental limitations in current approaches and outlining directions towards significantly enhanced operational performance. While existing FMs demonstrate remarkable performance on static benchmarks, their offline training paradigm, fixed temporal embeddings, and static cross-modal learned relationships limit their applicability to the dynamic, non-stationary nature of continuous EO data.

Three possible research directions have been identified to realize this vision. First, developing updating mechanisms via selective replay strategies and distillation-based approaches that update representations while maintaining consistency with past embeddings, without requiring full model retraining. Second, creating explicit temporal architectures to discriminate cyclical patterns from distributional shifts (e.g.,

via multi-scale temporal attention-based encoders) and integrating trend/shift detection modules for long-term changes, to support robust detection and adaptation. Third, implementing context-dependent cross-modal fusion that dynamically adjusts modality relevance over time based on reliability scores, current observations, and historical performance, rather than applying fixed fusion weights learned during pretraining.

The convergence of FMs, CL, and SML represents not merely an incremental improvement but a necessary paradigm shift for operational EO systems. As satellite constellations expand and data generation rates accelerate, the ability to continuously integrate new information while maintaining learned representations becomes essential for EO applications that demand scalable and interpretable solutions across diverse spatiotemporal contexts, ranging from disaster response to climate change monitoring. The research directions outlined in this paper provide a roadmap for the community to develop systems that are simultaneously adaptive, multimodal, and drift aware, characteristics essential for an ever-changing world.

Acknowledgements

The author acknowledges financial support from TEF through a PhD fellowship, during which this paper was refined. The author would also like to thank Prof. Emanuele Della Valle and his PhD Candidate Federico Giannini for their valuable scientific discussions, insightful feedback, careful proofreading, and continuous guidance throughout the development of this work.

Declaration on Generative AI

During the preparation of this work, the author used Generative AI tools to improve grammar and readability (ChatGPT, Gemini, and Claude) and to translate selected passages (DeepL). In addition, the author used Gemini to generate and/or edit figures. After using these tools, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, N. Díaz-Rodríguez, Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges, Information fusion 58 (2020) 52–68.

[2] A. Bifet, R. Gavalda, G. Holmes, B. Pfahringer, Machine learning for data streams: with practical examples in MOA, MIT press, 2023.

[3] A. Cossu, F. Giannini, G. Ziffer, A. Bernardo, A. Gepperth, E. Della Valle, B. Hammer, D. Bacciu, A practical guide to streaming continual learning, Neurocomputing 674 (2026) 132951. URL: https://www.sciencedirect.com/science/article/pii/S0925231226003486. doi:https://doi.org/10.1016/j.neucom.2026.132951.

[4] F. Giannini, G. Ziffer, A. Cossu, V. Lomonaco, Streaming continual learning for unified adaptive intelligence in dynamic environments, IEEE Intelligent Systems 39 (2024) 81–85.

[5] N. Gunasekara, B. Pfahringer, H. M. Gomes, A. Bifet, Survey on online streaming continual learning, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 6628–6637.

[6] L. Iovine, G. Ziffer, A. Proia, E. Della Valle, Towards streaming land use classification of images with temporal distribution shifts, in: ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, i6doc.com publ., Bruges (Belgium) and online, 2025. URL: http://www.i6doc.com/en/.

[7] A. Tsymbal, The problem of concept drift: definitions and related work, Computer Science Department, Trinity College Dublin 106 (2004) 58.

[8] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, F. Petitjean, Characterizing concept drift, Data Mining and Knowledge Discovery 30 (2016) 964–994.
[9] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, Z. Wang, S. Ebrahimi, H. Wang, Continual learning of large language models: A comprehensive survey, ACM Computing Surveys 58 (2025) 1–42.
[10] T. L. Hayes, G. P. Krishnan, M. Bazhenov, H. T. Siegelmann, T. J. Sejnowski, C. Kanan, Replay in deep learning: Current approaches and missing biological elements, Neural computation 33 (2021) 2908–2950.
[11] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural networks 113 (2019) 54–71.
[12] M. Bahri, A. Bifet, J. Gama, H. M. Gomes, S. Maniu, Data stream analysis: Foundations, major tasks and tools, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1405.
[13] R. Sebastião, J. M. Fernandes, Supporting the page-hinkley test with empirical mode decomposition for change detection, in: International symposium on methodologies for intelligent systems, Springer, 2017, pp. 492–498.
[14] L. Yang, N. Chen, J. Yue, Y. Liu, J. Ma, P. Ghamisi, A. Plaza, L. Fang, Survey of multimodal geospatial foundation models: Techniques, applications, and challenges, arXiv preprint arXiv:2510.22964 (2025).
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[16] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[17] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, S. Ermon, Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, Advances in Neural Information Processing Systems 35 (2022) 197–211.
[18] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, et al., Ringmo: A remote sensing foundation model with masked image modeling, IEEE Transactions on Geoscience and Remote Sensing 61 (2022) 1–22.
[19] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, J. Zhou, Remoteclip: A vision language foundation model for remote sensing, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–16.
[20] Z. Yu, C. Liu, L. Liu, Z. Shi, Z. Zou, Metaearth: A generative foundation model for global-scale remote sensing image generation, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[21] J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, R. Ramachandran, P. Fraccaro, T. Brunschwiler, G. Cavallaro, J. Bernabe-Moreno, N. Longépé, TerraMind: Large-Scale Generative Multimodality for Earth Observation, 2025. URL: http://arxiv.org/abs/2504.11171. doi:10.48550/arXiv.2504.11171, arXiv:2504.11171 [cs].
[22] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).
[23] A. Cossu, A. Carta, L. C. Passaro, V. Lomonaco, T. Tuytelaars, D. Bacciu, Continual pre-training mitigates forgetting in language and vision, Neural Networks 179 (2024) 106492.
[24] E. Fini, V. G. T. Da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, J. Mairal, Self-supervised models are continual learners, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 9621–9630.
[25] A. Gomez-Villa, B. Twardowski, L. Yu, A. D. Bagdanov, J. Van de Weijer, Continually learning self-supervised representations with projected functional regularization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3867–3877.
[26] S. Du, T. Li, Y. Yang, S.-J. Horng, Multivariate time series forecasting via attention-based encoder–decoder framework, Neurocomputing 388 (2020) 269–279.

[27] A. Bifet, R. Gavaldà, Learning from time-changing data with adaptive windowing, in: Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA, SIAM, 2007, pp. 443–448. URL: https://doi.org/10.1137/1.9781611972771.42. doi:10.1137/1.9781611972771.42.

[28] D. Yu, X. Zhang, Y. Chen, A. Liu, Y. Zhang, P. S. Yu, I. King, Recent advances of multimodal continual learning: A comprehensive survey, arXiv preprint arXiv:2410.05352 (2024).

[29] H. Wang, S. Zhou, Q. Wu, H. Li, F. Meng, L. Xu, H. Qiu, Confusion mixup regularized multimodal fusion network for continual egocentric activity recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3560–3569.

[30] G. Song, X. Tan, Real-world cross-modal retrieval via sequential learning, IEEE Transactions on Multimedia 23 (2020) 1708–1721.