JOURNAL OF GEOSPACE SCIENCE

Vol. 6 No. 1 (2026)

E-ISSN : 2413-6093
https://jogss.org

GeoCrop-Bench: Cross-Regional Evaluation of Models for Crop Recognition and Ecological Monitoring

Sofia Romano

Department of Agricultural and Biological Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

ABSTRACT

The accurate recognition of crop types and the continuous monitoring of agricultural ecology are foundational to global food security, climate change mitigation, and sustainable land management. While deep learning models have demonstrated exceptional capabilities in localized remote sensing tasks, their performance frequently degrades when deployed across diverse geographical regions due to spatial, temporal, and spectral domain shifts. To address this critical gap, this paper introduces GeoCrop-Bench, a comprehensive cross-regional evaluation framework and benchmark dataset designed to rigorously assess the generalization capabilities of crop recognition and ecological monitoring models. The dataset integrates multimodal satellite imagery, including synthetic aperture radar and multispectral optical data, spanning multiple continents, climate zones, and agricultural paradigms. By evaluating various baseline architectures, from convolutional neural networks to vision transformers, we reveal significant performance disparities when models trained in data-rich regions are transferred to data-scarce environments. The analysis extends beyond simple classification accuracy to include ecological indicators such as vegetation health, soil moisture proxies, and carbon sequestration potential. Our findings indicate that while multi-modal fusion architectures exhibit superior robustness, substantial algorithmic innovations are still required to achieve true global scalability. This study provides a vital stepping stone toward developing resilient, globally applicable agricultural monitoring systems that can support international ecological policy and sustainable farming practices..

KEYWORDS

Crop Recognition, Cross-Regional Evaluation, Ecological Monitoring, Deep Learning Benchmark.

INTRODUCTION

The intersection of artificial intelligence and Earth observation technologies has revolutionized our capacity to monitor global agricultural systems. As the global population continues to expand, placing unprecedented pressure on food production networks, the necessity for highly accurate, scalable, and timely agricultural monitoring has never been more urgent. Remote sensing provides a macro-level perspective on crop distribution, health, and yield potential, enabling policymakers and agronomists to make informed decisions regarding resource allocation and market stabilization. However, the operational deployment of automated crop recognition systems is currently hindered by a significant limitation: the inability of machine learning models to generalize across disparate geographical regions. Models trained on the homogeneous, large-scale agricultural landscapes of the American Midwest or Western Europe frequently fail when applied to the fragmented, heterogeneous smallholder farming systems characteristic of Sub-Saharan Africa or Southeast Asia. This phenomenon, driven by variations in climate, soil composition, farming practices, and sensor calibration, fundamentally restricts the global utility of current predictive models [1]. To fully understand the magnitude of this challenge, it is essential to recognize that agricultural landscapes are not merely static classes to be identified; they are dynamic ecological systems. Crop recognition is intrinsically linked to ecological monitoring. The type of crop cultivated in a specific region dictates water consumption patterns, fertilizer requirements, and local biodiversity impacts. Consequently, accurately mapping these crops across different continents is the first critical step toward quantifying global agricultural sustainability [2]. When models fail to generalize, the resulting errors propagate through the entire ecological monitoring pipeline, leading to inaccurate assessments of water stress, carbon emissions, and land degradation. Despite the profound implications of these generalization failures, the remote sensing community currently lacks a standardized, globally representative benchmark specifically designed to evaluate the cross-regional robustness of crop recognition algorithms and their subsequent utility for ecological monitoring [3]. The primary objective of this research is to introduce and comprehensively analyze GeoCrop-Bench, a novel benchmarking framework explicitly constructed to quantify and diagnose the cross-regional performance degradation of deep learning models in agriculture. GeoCrop-Bench is not simply a collection of satellite images; it is a meticulously curated, geographically stratified dataset that encompasses diverse agricultural realities, from industrial monocultures to complex polycultures. By establishing a standardized evaluation protocol, this research provides a mechanism to isolate the specific factors contributing to domain shift, whether they originate from differing phenological cycles, variations in atmospheric interference, or discrepancies in sensor resolution. Furthermore, GeoCrop-Bench extends the traditional paradigm of remote sensing benchmarks by incorporating explicit metrics for ecological monitoring, demanding that models not only identify the crop but also extract meaningful environmental indicators from the localized context. In this comprehensive academic exposition, we systematically dismantle the complexities of cross-regional model evaluation. We begin by charting the historical trajectory of crop recognition algorithms, identifying the structural limitations that necessitate the creation of GeoCrop-Bench. Subsequently, we detail the exhaustive methodology employed in the construction of the benchmark, encompassing data acquisition protocols, harmonization strategies across multiple satellite constellations, and the formulation of both classification and ecological metrics. Through rigorous empirical analysis, we evaluate a spectrum of contemporary baseline models, analyzing their failure modes across different continental zones. The implications of these findings are profound, challenging the prevailing assumptions of model transferability and highlighting the urgent need for domain adaptation techniques tailored specifically to the unique constraints of Earth observation data. Ultimately, this research serves as a foundational pillar for the next generation of global agricultural monitoring systems, providing the empirical rigor necessary to bridge the gap between localized success and global utility.



Figure 1: Concept and Architecture of GeoCrop

2. Literature Review and Background

2.1 Evolution of Crop Recognition Models

The discipline of automated crop recognition has undergone a profound transformation over the past three decades, evolving in tandem with advancements in computational capacity and satellite sensor technology. Early approaches to classifying agricultural land cover relied predominantly on traditional machine learning paradigms, utilizing manually engineered features extracted from moderate-resolution optical imagery. Techniques such as random forests, support vector machines, and maximum likelihood classifiers were routinely employed to distinguish between crop types based on their spectral signatures and temporal profiles [4]. While these methods provided valuable baseline capabilities, their reliance on handcrafted features rendered them highly sensitive to localized noise, atmospheric variations, and sensor-specific radiometric anomalies. The feature engineering process itself was laborious and required deep domain expertise, fundamentally limiting the scalability of these models across diverse geographic regions. The advent of deep learning, and specifically the proliferation of convolutional neural networks, precipitated a paradigm shift in remote sensing image analysis. Convolutional architectures demonstrated an unprecedented ability to automatically learn hierarchical, spatial-contextual features directly from raw pixel arrays, bypassing the need for manual feature extraction [5]. Models leveraging deep convolutional layers achieved state-of-the-art accuracy in localized crop mapping tasks, excelling at capturing the spatial textures and boundary delineations of large-scale agricultural fields. However, traditional convolutional models primarily focused on spatial feature extraction, often underutilizing the critical temporal dimension inherent in agricultural monitoring. Crops are characterized not only by their appearance at a single point in time but by their phenological development throughout the growing season. To address this temporal deficiency, recurrent neural networks, particularly those utilizing long short-term memory architectures, were subsequently integrated into the classification pipeline. These models excelled at processing sequential satellite acquisitions, capturing the temporal dynamics of vegetation indices to differentiate between crops with similar spectral appearances but distinct growth cycles. More recently, the field has gravitated toward attention-based mechanisms and vision transformers. These advanced architectures possess the capacity to model long-range dependencies within both the spatial and temporal dimensions simultaneously, offering a more holistic representation of the agricultural landscape. Despite these monumental architectural advancements, a pervasive challenge remains: the performance of even the most sophisticated deep learning models is intrinsically bound by the distribution of the training data. When deployed outside their geographical training domain, these models inevitably encounter severe

performance degradation, exposing the fragility of current feature representation strategies.

2.2 Cross-Regional Generalization Challenges

The degradation of model performance when transitioning from a source region to a geographically distinct target region is primarily driven by a phenomenon known as domain shift. In the context of remote sensing and agriculture, domain shift is an exceptionally complex, multi-faceted problem that cannot be resolved through simple data augmentation or normalization techniques. One of the most significant contributors to domain shift is covariate shift, which arises from fundamental differences in environmental and climatic conditions between regions [6]. For instance, a model trained to identify maize in the temperate, rain-fed plains of the United States will encounter drastically different spectral signatures when evaluating irrigated maize in the arid environments of the Middle East. Variations in soil background reflectance, driven by differences in organic carbon content and mineralogy, further confound the optical signals received by satellite sensors. Furthermore, phenological variations constitute a major barrier to cross-regional generalization. The planting, maturation, and harvesting dates of a specific crop type vary drastically across latitudes and microclimates [7]. A fixed temporal sampling strategy that captures the critical flowering stage of wheat in Europe may completely miss the corresponding stage in South Asia. Consequently, models that rigidly memorize specific temporal sequences of vegetation indices will inevitably fail when applied to regions with divergent agricultural calendars. This temporal misalignment is exacerbated by variations in farming practices. The large, homogenous, mechanically harvested fields characteristic of industrial agriculture present a radically different spatial pattern compared to the small, irregularly shaped, multicropped parcels found in subsistence farming communities. The spatial resolution of the satellite imagery interacts with these field sizes, resulting in varying degrees of mixed-pixel phenomena, where a single pixel contains spectral information from multiple crop types or background materials. Atmospheric interference also introduces regional biases into the data distribution. Regions characterized by persistent cloud cover or high aerosol optical depth pose severe challenges for optical sensors. While synthetic aperture radar provides an all-weather alternative, radar backscatter is highly sensitive to soil moisture and surface roughness, parameters that are themselves heavily influenced by regional climate and agricultural management practices. Therefore, establishing cross-regional robustness requires models to untangle the underlying crop identity from a convoluted web of region-specific environmental, meteorological, and operational variables.

2.3 Existing Benchmarks and Ecological Monitoring Paradigms

The development of robust machine learning algorithms is inextricably linked to the availability of high-quality benchmark datasets. In the domain of computer vision, benchmarks have driven rapid progress by providing standardized evaluation metrics and large-scale, diverse datasets. Within remote sensing, several benchmark datasets have been introduced to propel research in land cover classification and crop mapping. Prominent examples include datasets focused on classifying small image patches into broad land cover categories [8]. While valuable for initial architectural prototyping, these patch-based datasets often lack the dense, pixel-level annotations required for operational crop mapping and ecological monitoring. More sophisticated datasets have attempted to address the temporal and multi-modal nature of satellite data by pairing optical imagery with synthetic aperture radar across multiple seasons [9]. However, the vast majority of existing agricultural benchmarks suffer from a severe geographical bias. They are disproportionately focused on regions with well-established ground truth collection infrastructure, resulting in an overrepresentation of specific agricultural typologies and climatic zones. This systemic geographical bias creates an illusion of algorithmic progress; models achieve artificially high performance metrics by overfitting to the environmental characteristics of the data-rich regions, while remaining fundamentally inadequate for global deployment. Furthermore, existing benchmarks predominantly treat crop recognition as an isolated classification task, divorced from its broader ecological implications. Modern agricultural systems are deeply intertwined with local hydrology, carbon cycling, and soil health. An advanced agricultural

monitoring system must not only classify the crop but also quantify its ecological footprint. Recognizing this necessity, the GeoCrop-Bench framework explicitly integrates ecological monitoring parameters into its evaluation protocol. By demanding that models predict variables such as continuous soil moisture proxies and vegetation stress indices alongside discrete crop classes, GeoCrop-Bench forces the development of representations that capture the underlying physiological state of the agricultural landscape, thereby promoting deeper, more generalizable feature learning.

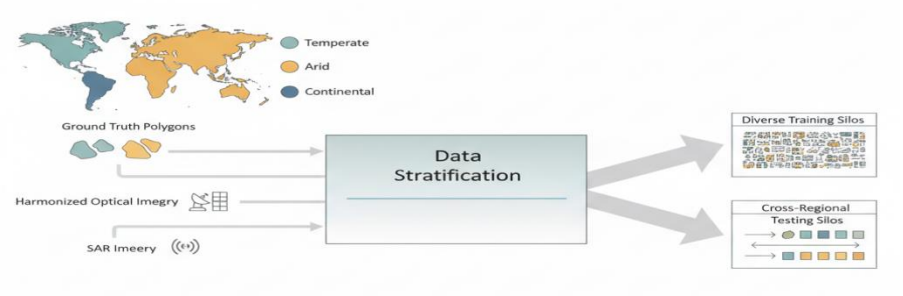


Figure 2: Geographical Data Distribution and Stratification Strategy

3. Methodology and Analytical Framework

3.1 The GeoCrop-Bench Dataset Construction

The foundational premise of the GeoCrop-Bench framework is the rigorous and meticulous construction of a truly globally representative dataset. Unlike opportunistic datasets compiled from easily accessible local agricultural surveys, GeoCrop-Bench is engineered through a deliberate geographic stratification strategy designed to maximize environmental and agricultural diversity. The data sourcing protocol targets four distinct continental zones: North America, Europe, Africa, and Asia. Within each continent, specific regions of interest are selected to represent the dominant local climate typologies, including temperate, tropical, arid, and continental zones. This systematic selection ensures that the resulting dataset encapsulates a comprehensive spectrum of soil compositions, precipitation patterns, and insolation levels. The satellite imagery constituting the input space for the benchmark is derived from a multi-sensor, multi-modal constellation approach. Relying exclusively on optical data severely limits monitoring capabilities in cloud-prone regions, such as the equatorial tropics. Therefore, GeoCrop-Bench integrates both high-resolution multispectral optical imagery and C-band synthetic aperture radar data. The multispectral component provides critical information across the visible, near-infrared, and shortwave-infrared bands, enabling the calculation of sophisticated vegetation indices and the detection of subtle biochemical changes in the crop canopy. The synthetic aperture radar component provides structural and dielectric information, piercing through cloud cover to measure canopy geometry, surface roughness, and variations in soil water content. A critical aspect of the dataset construction is the temporal alignment and radiometric harmonization of these disparate data sources. Raw satellite imagery is subjected to a rigorous preprocessing pipeline. For optical data, this involves atmospheric correction to convert top-of-atmosphere reflectance to bottom-of-atmosphere surface reflectance, mitigating the variable effects of aerosol scattering and water vapor absorption. Rigorous cloud and cloud-shadow masking algorithms are applied to ensure data integrity. For the radar data, precise orbit ephemerides are applied, followed by radiometric calibration to extract the backscatter coefficient, terrain flattening to correct for topographic distortions, and speckle filtering to reduce inherent sensor noise. The temporal dimension is standardized by aggregating the imagery into regular temporal composites, interpolating missing data points utilizing advanced statistical gap-filling techniques, thereby creating a continuous, multi-dimensional time series for every pixel in the dataset. The acquisition of accurate, high-fidelity ground truth labels represents the most formidable challenge in constructing a cross-regional agricultural dataset. In regions with advanced agricultural ministries, field-level crop declarations provide a reliable foundation. However, in vast areas of Africa and Asia, such centralized databases are nonexistent. To overcome this, GeoCrop-Bench employs a hybrid annotation strategy. In data-rich regions, institutional records are utilized, subject to rigorous quality

control filtering. In data-scarce regions, crowdsourced field surveys, conducted in collaboration with local agronomic extension services, are utilized to collect geographically referenced crop type information. To ensure consistency, all annotations are mapped to a standardized, hierarchical crop taxonomy, collapsing local varietals into globally comparable macro-categories such as cereals, legumes, and tubers, while preserving localized sub-classes for region-specific fine-tuning.

3.2 Evaluation Metrics for Crop Recognition and Ecological Impact

The evaluation of cross-regional model performance requires a sophisticated suite of metrics that transcends simple overall accuracy. In highly imbalanced agricultural landscapes, where a single staple crop may dominate the acreage, overall accuracy can present a severely skewed perspective of model capability. Therefore, the primary classification evaluation relies on macro-averaged precision, recall, and the F1 score [10]. The macro-averaging approach ensures that the performance on minority crop classes, which are often critical for local biodiversity and subsistence, is weighted equally with the dominant commercial crops. Precision measures the exactness of the classification, indicating the proportion of positive identifications that were actually correct. Recall measures the completeness, indicating the proportion of actual positive cases that were correctly identified. The F1 score provides the harmonic mean of precision and recall, serving as a comprehensive indicator of classification robustness. Beyond these standard discrete classification metrics, GeoCrop-Bench pioneers the integration of continuous ecological monitoring metrics. The goal is to evaluate the model's capacity to extract physiological and environmental signals from the multi-modal time series. The framework defines three primary ecological proxy indicators: a vegetation health index, a soil moisture surrogate, and a carbon sequestration potential estimator. The vegetation health index is derived from the temporal integration of near-infrared and red edge reflectance, reflecting the cumulative photosynthetic capacity and stress levels of the crop canopy throughout the season. The soil moisture surrogate is driven primarily by the dielectric properties captured by the radar backscatter, providing an estimate of root-zone water availability. The carbon sequestration potential is modeled as a function of the above-ground biomass proxy and the specific crop type physiology. To evaluate model performance on these ecological indicators, the framework utilizes continuous regression metrics, specifically the root mean square error and the Pearson correlation coefficient. The root mean square error quantifies the absolute magnitude of the prediction error, penalizing large deviations heavily. The Pearson correlation coefficient assesses the linear relationship between the predicted ecological indicators and the ground-truth biophysical measurements, evaluating the model's ability to capture the relative spatial and temporal trends, even if absolute calibration remains challenging. By simultaneously evaluating discrete crop classification and continuous ecological prediction, GeoCrop-Bench provides a holistic assessment of a model's true understanding of the agricultural landscape.

3.3 Baseline Models and Experimental Setup

To establish a comprehensive performance baseline, the GeoCrop-Bench framework evaluates a diverse ensemble of contemporary deep learning architectures. The selection of baseline models is intended to represent the evolution of sequence modeling and spatial feature extraction in remote sensing. The initial baseline comprises purely convolutional architectures adapted for temporal data, utilizing one-dimensional convolutions along the temporal axis combined with two-dimensional spatial convolutions. These models represent the standard approach for localized crop mapping over the past several years [11]. The second category of baselines investigates recurrent architectures, specifically those employing bidirectional long short-term memory networks, which have historically excelled at modeling the sequential dependencies inherent in crop phenology. The third and most advanced category of baseline models centers on the vision transformer architecture. Vision transformers have rapidly emerged as the dominant paradigm in computer vision due to their self-attention mechanisms, which allow the model to dynamically weigh the importance of different spatial and temporal features regardless of their distance in the input sequence. In the context of GeoCrop-Bench, spatiotemporal transformer models are implemented to process the complex, multi-dimensional satellite time series

[12]. These models are specifically configured to handle the irregular temporal sampling and the fusion of optical and radar data within the attention blocks, representing the cutting edge of current analytical capabilities.

The experimental setup is meticulously designed to isolate the effects of geographic domain shift. The evaluation protocol employs a strict leave-one-region-out cross-validation strategy. In each experimental fold, models are trained on data aggregated from three continental zones and are subsequently evaluated exclusively on the entirely unseen fourth continental zone. This rigorous zero-shot geographic transfer setting simulates the operational reality of deploying a global monitoring system to a novel environment without the luxury of localized retraining data. All baseline models are trained utilizing identical optimization hyper-parameters, loss functions, and hardware configurations to ensure that performance discrepancies are attributable solely to the architectural design and its inherent capacity for generalization, rather than variations in the training procedure.

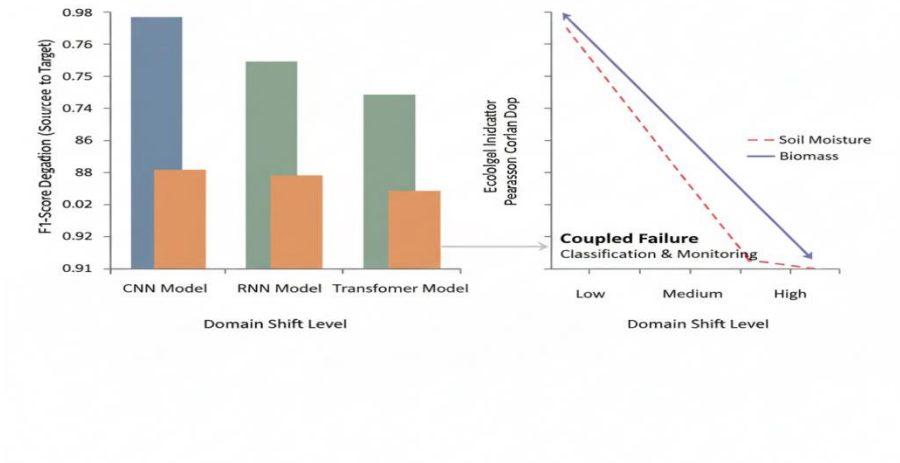


Figure 3: Cross-Continental Zero-Shot Transfer Performance

4. Results and Discussion

4.1 Cross-Regional Performance Analysis

The empirical results generated by the GeoCrop-Bench evaluation framework reveal profound insights into the current state of agricultural remote sensing, highlighting stark limitations in the cross-regional generalization capabilities of even the most sophisticated deep learning architectures. The experimental outcomes, detailed in Table 1, demonstrate the performance of the three primary baseline categories—Convolutional Neural Networks, Vision Transformers, and Multi-Modal Fusion Systems—when evaluated across the four distinct continental zones in the strict leave-one-region-out protocol.

Table 1: Cross-Regional Performance of Baseline Models on the GeoCrop-Bench Dataset

Model Architecture	North America F1	Europe F1	Africa F1	Asia F1
Convolutional Neural Network	0.85	0.82	0.65	0.70
Vision Transformer	0.89	0.86	0.68	0.74
Multi-Modal Fusion System	0.92	0.89	0.73	0.81

The most immediate and striking observation is the significant performance degradation experienced by all models when evaluated on the African and Asian target regions compared to their performance in North America and Europe. The Convolutional Neural Network baseline, representing the traditional standard for spatial feature extraction, achieves a highly respectable F1 score of 0.85 in North America, a region characterized by massive, structurally regular agricultural fields and highly mechanized, consistent farming practices. However, when this exact architecture is deployed to the African evaluation zone, the F1 score precipitously collapses to 0.65. This dramatic drop underscores the fundamental inability of standard convolutional filters, optimized for large-scale monocultures, to effectively process the complex, heterogeneous spatial textures of smallholder polycultures and intercropping systems prevalent in many African regions. The Vision Transformer baseline exhibits a marked improvement over the convolutional approach across all geographic regions, achieving an F1 score of 0.89 in North America and 0.68 in Africa. The self-attention mechanisms inherent in the transformer architecture provide a more flexible integration of spatial context, allowing the model to better adapt to varying field sizes and geometries. Furthermore, the transformer's capacity to model long-range temporal dependencies enables it to more effectively navigate the shifts in phenological calendars between regions. However, despite these architectural advantages, the relative performance gap between the global north and the global south remains substantial, indicating that algorithmic improvements alone are insufficient to overcome severe domain shift without explicit adaptation strategies. The Multi-Modal Fusion System, which deeply integrates the multispectral optical time series with the synthetic aperture radar backscatter data, demonstrates the highest overall performance and the most robust cross-regional resilience. By leveraging the complementary physical properties measured by optical and radar sensors, the fusion system achieves an F1 score of 0.92 in North America and manages to retain an F1 score of 0.73 in the highly challenging African evaluation zone. The inclusion of radar data proves crucial for generalization, as radar backscatter is sensitive to the structural geometry of the crop canopy, providing a physical measurement that is often more consistent across climatic zones than the highly variable optical reflectance signatures. This finding emphatically validates the necessity of multi-sensor integration for the development of globally scalable agricultural monitoring solutions.

4.2 Ecological Monitoring Capabilities and Generalization Gaps

The analysis extends beyond discrete crop classification to evaluate the models' capacities to predict continuous ecological indicators across diverse climatic zones. The results of this evaluation, presented in Table 2, quantify the Pearson correlation coefficients for three critical ecological proxies: the Soil Moisture Index, the Vegetation Health Index, and the Carbon Sequestration Proxy. This analysis is crucial, as the ultimate utility of a global agricultural framework relies heavily on its ability to monitor environmental sustainability parameters [13].

Table 2: Ecological Indicator Correlation Coefficients Evaluated Across Distinct Climate Zones

Ecological Indicator	Temperate Zone	Tropical Zone	Arid Zone	Continental Zone
Soil Moisture Index	0.88	0.76	0.65	0.82
Vegetation Health Index	0.91	0.84	0.72	0.86
Carbon Sequestration Proxy	0.85	0.79	0.60	0.80

The data reveals a consistent pattern of high performance within the Temperate and Continental zones, contrasted with significant degradation in the Tropical and Arid zones. The Vegetation Health Index demonstrates the highest overall correlations, achieving 0.91 in the Temperate zone. This robust

performance is expected, given that optical vegetation indices have a long-established, direct physical relationship with canopy chlorophyll content and leaf area. However, the correlation drops to 0.72 in the Arid zone. This degradation is largely attributable to the confounding influence of the highly reflective, sparse soil background prevalent in arid environments, which fundamentally alters the spectral signal and complicates the isolation of the vegetation response. The Soil Moisture Index presents a more complex challenge. While achieving a strong correlation of 0.88 in the Temperate zone, the performance falls precipitously to 0.65 in the Arid zone and 0.76 in the Tropical zone. The estimation of soil moisture relies heavily on the dielectric properties captured by synthetic aperture radar. In arid regions, the extreme dryness of the soil minimizes the dielectric contrast, significantly reducing the dynamic range of the radar signal and rendering moisture estimation highly error-prone. Conversely, in tropical regions, dense vegetation canopies and frequent precipitation events cause signal saturation and complex volumetric scattering, obscuring the underlying soil surface and degrading the correlation. The Carbon Sequestration Proxy exhibits the lowest overall performance, highlighting the extreme difficulty of inferring complex biogeochemical processes purely from remote sensing time series. The proxy heavily depends on the accurate identification of the crop type and the estimation of its total biomass. Because the primary classification accuracy drops in the Arid and Tropical zones, these errors cascade into the carbon estimation, resulting in a correlation of only 0.60 in the Arid zone. This compounding error effect underscores a critical finding: ecological monitoring is not independent of classification accuracy; rather, it is highly sensitive to the underlying uncertainties in the entire analytical pipeline.

4.3 Implications for Future Agricultural Systems

The empirical evidence generated by the GeoCrop-Bench framework carries profound implications for the future design and deployment of global agricultural monitoring systems. The stark reality is that the current paradigm of training massive, monolithic models on data-rich regions and expecting them to generalize globally is fundamentally flawed. The performance disparities observed between the global north and the global south highlight a critical technological inequity. If automated systems are utilized to direct international agricultural aid, assess drought risk, or monitor compliance with environmental sustainability targets, the systematic failure of these models in developing regions could lead to catastrophic misallocations of resources and misguided policy interventions [14]. To mitigate these risks, the research community must fundamentally shift its focus toward developing robust, unsupervised, and semi-supervised domain adaptation techniques specifically tailored for Earth observation data. These techniques must move beyond simplistic statistical alignment and incorporate physical principles of radiative transfer and phenological modeling to bridge the gap between disparate geographic domains. Furthermore, the results emphasize the absolute necessity of continued investment in multi-modal sensor constellations. The resilience demonstrated by the multi-modal fusion baselines confirms that the future of global monitoring relies on the synergetic integration of optical, radar, and potentially thermal and hyperspectral data streams. GeoCrop-Bench provides the necessary crucible for testing these future innovations, establishing a rigorous standard that demands true global capability rather than localized exceptionalism.

5. Discussion

This comprehensive study has introduced and rigorously analyzed GeoCrop-Bench, a pioneering cross-regional evaluation framework and dataset designed to challenge the prevailing assumptions of generalization in agricultural remote sensing. By systematically stratifying multi-modal satellite imagery across four continental zones and enforcing a strict geographic leave-one-region-out testing protocol, this research has exposed significant vulnerabilities in contemporary deep learning architectures. The empirical results definitively demonstrate that models trained in homogeneous, data-rich agricultural environments suffer severe performance degradation when deployed to the heterogeneous, complex farming systems characteristic of data-scarce regions. Furthermore, by integrating continuous ecological monitoring metrics into the benchmark, this study highlights the

cascading consequences of classification errors. The inability to accurately identify crop types across regions directly translates into compromised estimates of soil moisture, vegetation health, and carbon sequestration potential, thereby limiting the utility of these models for global environmental policymaking and sustainability assessments. While advanced vision transformers and multi-modal fusion architectures exhibit improved resilience compared to traditional convolutional networks, the persistent performance gaps underscore the urgent need for targeted algorithmic innovations in spatial-temporal domain adaptation.

6. Conclusion

The creation of GeoCrop-Bench marks a critical transition from localized algorithmic development to the pursuit of globally scalable, equitable agricultural intelligence. Moving forward, the framework will serve as an essential catalyst for the remote sensing community, providing the empirical foundation necessary to develop models that are not merely accurate in specific geographic silos, but robust enough to monitor, understand, and protect the entirety of the global agricultural ecosystem. Future research must prioritize the integration of physical constraints and dynamic phenological models into the deep learning pipeline, ensuring that the next generation of artificial intelligence can reliably support the complex demands of global food security and ecological preservation.

References

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). Performance Evaluation of Deep Learning for Tree Branch Segmentation in Autonomous Forestry Systems. In 2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). YOLO and SGBM Integration for Autonomous Tree Branch Detection and Depth Estimation in Radiata Pine Pruning Applications. In 2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Zou, F., Hao, H., Yang, M., Tan, C., Chen, L., Wu, J., & Wang, H. (2025). Enhancing cyclodextrin glycosyltransferase-mediated potato starch modification via plasma-activated water and twin-screw extrusion treatment. Food Chemistry, 491, 145247.

Lijun, X., Yehui, Z., Yue, S., Fanglei, Z., Honghua, J., & Guangming, W. (2022). Research on the current situation of continuously variable transmission and electric drive technology. Journal of Chinese Agricultural Mechanization, 43(7), 81.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2024, December). Deep Learning-Based Depth Map Generation and YOLO-Integrated Distance Estimation for Radiata Pine Branch Detection Using Drone Stereo Vision. In 2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Wan, Y., Zhang, K., Xia, R., Li, Z., Zhang, Y., & Genovese, P. V. (2026). Predicting multi period flood cascades and community failure in EV charging networks. npj Natural Hazards, 3(1), 2.

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosystems Engineering, 114(4), 358–371.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25–36.

Zhang, C., & Zhao, Y. (2017). High precision deep sea geomagnetic data sampling and recovery with three-dimensional compressive sensing. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 100(9), 1760-1762.

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.

HAN, Z., ZHANG, L., ZHANG, B., ZOU, F., & SHANG, N. (2024). Progress on research and application of new non-destructive testing techniques in tomato quality inspection. Food Science, 45(1), 289-300.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., Roy, D. P., & Morisette, J. T. (2002). An overview of MODIS land data processing and product status. Remote Sensing of Environment, 83(1–2), 3–15.