AG-STELLA: Spatio-Temporal Learning for Water-related Agricultural Land Use Activity Mapping with AlphaEarth

Nibir Chandra Mandal[1,2], **Oishee Bintey Hoque**[1,2], **Kyle Luong**[1], **Samarth Swarup**[2], **Kirti Rajagopalan**[3], **Mandy L Wilson**[2], **Abhijin Adiga**[2] and **Madhav Marathe**[1,2]
[1]Dept. of Computer Science, University of Virginia
[2]Biocomplexity Institute, University of Virginia
[3]Dept. of Biological Systems Engineering, Washington State University
{wyr6fx, gza5dr, auj7tx, swarup, alw4ey, abhijin, marathe}@virginia.edu

Abstract

Accurate mapping of agricultural land use activity, particularly long-term transition from cropland to pasture and short-term transition between cropland to fallow land, is essential for sustainable water management, drought response, and food-system resilience which directly supports United Nations Sustainable Development Goals (SDG-2 and SDG-8). However, reliable agricultural land use activity mapping is challenging due to spectral ambiguity, temporal irregularities, severe class imbalance, and limited generalization across agricultural regions. In this work, we propose AG-STELLA, a knowledge guided spatio-temporal model that (i) captures temporal changes of agricultural lands using pretrained spatio-temporal transformers; (ii) integrates geospatial context using AlphaEarth embedding; (iii) introduces a temporal transition latent space with temporal consistency constraints; (iv) employs guidance through hydroclimatic consistency; and (v) uses a land use-aware gated decoder to improve robustness across regions. Through experimentation across three water-stressed U.S. states, we show consistent gains over baseline vision and foundation models, achieving up to 27% F1-score improvement for pasture (minority class) and 16% overall. We further show the robustness across heterogeneous regions through cross-state transfer learning, where AG-STELLA consistently outperforms foundation model baselines and achieve up to 82.3% F1-score for fallow land with a 9.6% improvement over the best foundation model. Code: https://github.com/Nibir088/AG-STELLA

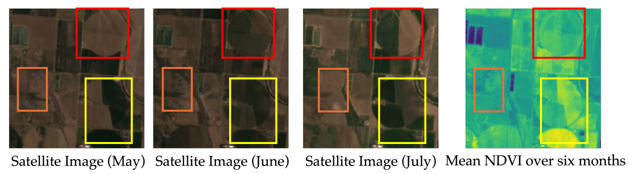

Figure 1: Three-month RGB imagery reveals distinct temporal patterns of agricultural land use activity across pasture (red), fallow (orange), and cropland (yellow). Fallow land remain largely unchanged over time, whereas cropland exhibits pronounced seasonal variation. The NDVI map averaged from April to October highlights higher vegetation values for cropland and lower values for fallow land.

1 Introduction

Agricultural land use activity refers to the temporal transitions of farmland among cropland, pasture, and fallow states, reflecting farmer responses to climate variability [Norton *et al.*, 2021], water availability [Tortajada *et al.*, 2017], and economic conditions [Żarczyński *et al.*, 2023]. In water stressed regions, these transitions increasingly emerge as adaptive responses to water scarcity and drought [Wang, 2023], serving as early indicators of irrigation constraints, groundwater depletion, and agricultural risk [Satapathy *et al.*, 2024]. Moreover, evidence from highly crop-intensive regions shows that short-term fallowing is commonly used to reduce groundwater withdrawals during drought [Chen *et al.*, 2023; Eslamifar *et al.*, 2024], while longer-term adaptation often involves transitions from irrigated cropland to lower-intensity pasture or grass-based systems under persistent water stress [Wang, 2023]. These transition reflects both water conservation responses [Mirzaei and Garcia, 2024] and structural shifts in agricultural production systems [Greiner *et al.*, 2021]. Accurate and timely mapping of land use activity is therefore essential for crop yield forecasting, drought response, supply-chain planning, and long-term sustainability assessments—particularly as fields increasingly rotate between productive and low-intensity states within short time horizons [Liu *et al.*, 2025; Song *et al.*, 2022]. For instance, drought relief decisions often require evidence of agricultural hardship beyond reduced water supply; quantifying short-term transitions from cropland to fallow or pasture can directly support faster and equitable drought response [WA State Legislature, 2018]. By enabling transparent assessment of agricultural stress, adaptive land management, land use activity mapping informs water allocation, safeguards ecosystem services, and supports rural livelihoods, thereby directly aligning agricultural land use activity mapping with the United Nations Sustainable Development Goals SDG 2 (Zero Hunger) and SDG 8 (Decent Work and Economic Growth) [WWAP, 2019].

Agricultural land use activity mapping faces several technical challenges. (i) Spectral ambiguity (as shown in Figure 1): Static satellite imagery suffers from severe spectral ambiguity, as fallow land, dormant pasture, and harvested cropland often appear visually indistinguishable [Morais *et al.*, 2024]. Classification algorithms routinely conflate these two vital states due to low spectral disparity across most bands [Bak *et al.*, 2025; Lei *et al.*, 2024]. Existing methods rely heavily on vegetation indices such as Normalized Difference Vegetation Index (NDVI), yet vegetation estimates are highly sensitive to data-quality issues, such as cloud occlusion, snow cover, atmospheric effects, and sensor noise [Hoque *et al.*, 2025]. Moreover, in regions with high crop diversity, overlapping crop temporal and spectral signatures amplify the ambiguity (e.g., weeds in a fallowed field vs fallowed of another crop) [Noorazar *et al.*, 2025]. (ii) Temporal sparsity: Temporal modeling is further complicated by irregular satellite revisit schedules and missing observations, which produce incomplete and noisy multi-month time series and obscure key temporal cues such as green-up when plants start to grow, and senescence when plants get mature and stop growing or start yellowing. (iii) Context dependence: Beyond temporal challenges, land use decisions are strongly context-dependent: visually similar patterns may arise from fundamentally different drivers, such as water scarcity–induced fallowing in arid regions versus soil-restoration practices in water-sufficient areas. (iv) Limited annotated data with severe class imbalances: Supervised learning approaches additionally face high annotation costs, data scarcity, and severe class imbalances, particularly for a minority activity class such as fallow land (e.g., only 6.9% for UT) and pasture land (e.g., 7.8% for CO). While recent foundation models offer strong representations, they remain insufficient for agricultural land use activity mapping. (v) Limitations of existing models. State-of-the-art models like Prithvi [Li *et al.*, 2023], though pretrained on large-scale remote sensing data and robust to noise and missing observations, lacks explicit agronomic reasoning needed to distinguish fallow land from pasture. Recent geospatial embeddings such as AlphaEarth [Brown *et al.*, 2025] introduce valuable spatial and geolocational context. However, these representations are typically provided at an annual temporal scale which limits their ability to capture fine-grained intra-seasonal management dynamics such as crop rotation, green-up, and senescence. As a result, existing methods continue to conflate visually similar yet agronomically and societally distinct land use states.

Contributions. To address the challenges, we propose **AG-STELLA**, a knowledge guided spatio-temporal model for agricultural land use activity mapping. Our main contributions are as follows:

- **End-to-end foundation-model–driven spatio-temporal model.** We introduce AG-STELLA, an end-to-end training model that jointly leverages Prithvi for multi-month spatio-temporal representation learning and AlphaEarth for geospatial context encoding. By coupling Prithvi-based spatio-temporal transformer representations with AlphaEarth embedding conditioned feature modulation and land use aware segmentation, AG-STELLA learns the representation that encodes intra-seasonal land use activity by emphasizing temporal changes in feature activations rather than static spectral intensity.
- **Transition-aware temporal representation.** We propose a *Temporal Transition Latent Space* that captures intra-seasonal dynamics from pretrained transformer tokens, guided by vegetation index variation to emphasize land use activity rather than static appearance.
- **Knowledge guidance through geo-spatial context and hydroclimatic constraints.** We incorporate geospatial context using AlphaEarth embedding to condition the model prediction and introduce hydroclimatic consistency constraints by imposing a class-wise ET–precipitation correlation constraint to enable context-aware predictions that generalize across heterogeneous environmental and climatic regimes.
- **Land Use-Aware Spatial Decoder.** We introduce a land use–aware spatial gating module that conditions the decoder on coarse land use priors to modulate spatial feature responses. By restricting feature activation to land use consistent regions, the model suppresses background interference and prioritizes land use activity relevant regions, resulting in more precise land use delineation across heterogeneous agricultural landscapes.
- **Extensive evaluation in state-specific and cross-state setting.** We conduct a comprehensive cross-state and in-state evaluation across Colorado, Washington, and Utah to assess generalization under diverse agricultural and cropping regimes. We evaluate across three U.S. states (Colorado, Washington, and Utah) and demonstrate consistent performance gains over state-of-the-art baselines and improved transferability across the U.S. states. AG-STELLA outperforms Prithvi and AlphaEarth with up to **16.0%** higher macro accuracy and **27.0%** higher F1-scores. Beyond performance metrics, we analyze crop-specific sources of systematic land use mapping errors across regions.

2 Related Work

Remote Sensing for Mapping Agricultural Land and Infrastructure. Deep learning has been widely applied in agricultural remote sensing for crop classification, field boundary detection, and irrigation mapping [Weiss *et al.*, 2020; Paolini *et al.*, 2022]. Convolutional Neural Network (CNN)-based models have been used for invasive species classification [Hung *et al.*, 2014], mixed-crop segmentation [Mortensen *et al.*, 2016], and weed detection [Di Cicco *et al.*, 2017], while attention-based architectures and channel-wise feature selection improve performance in complex landscapes [Cheng *et al.*, 2021]. However, both CNNs and Vision Transformers (ViTs) struggle to distinguish irrigation and land use classes with strong spectral similarity, and multi-stream fusion of spectral indices (NDVI, Normalized Difference Water Index (NDWI), and Normalized Difference Turbidity Index (NDTI)) improves robustness but still requires large labeled datasets for reliable generalization [Touvron *et al.*, 2021].

Monitoring Agricultural Assets and Land using Remote Sensing. Remote sensing is widely used to analyze agricultural dynamics and how climate, water, and socioeconomic

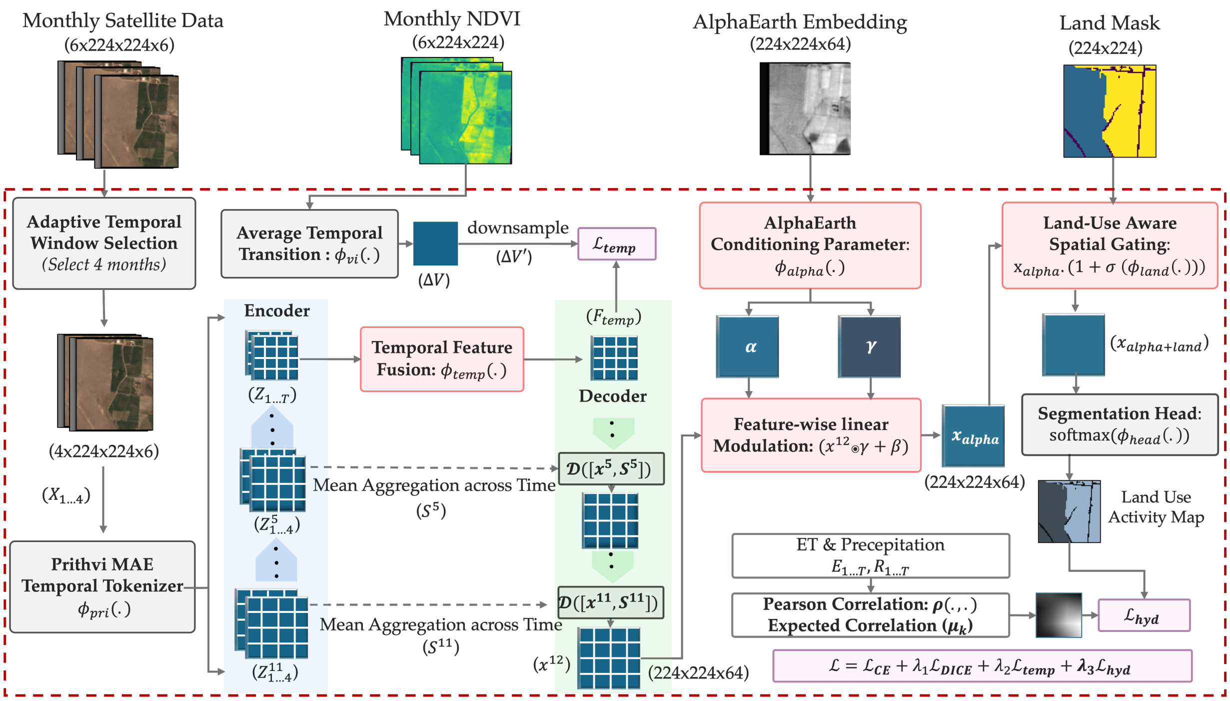

Figure 2: Architecture of AG-STELLA consists of three stages: (left) spatio-temporal encoding with adaptive window selection and Prithvi-based tokenization, (center) temporal transition latent space construction via feature fusion, and (right) context-conditioned decoding using AlphaEarth embeddings, land use–aware spatial gating, and hydroclimatic consistency constraints (bottom).

factors influence land-use decisions [Norton *et al.*, 2021]. Deep time-series models on Landsat, MODIS, and Sentinel enable large-scale detection of fallowed and abandoned cropland [Liu *et al.*, 2025; Koley and Chockalingam, 2025], while vegetation indices and learning-based models support pasture identification and condition monitoring [Christofi *et al.*, 2025; Rosanna *et al.*, 2025]. Yet, single-modality cues struggle under spectral ambiguity between fallow and pasture.

Semantic Segmentation Models. CNN-based u-net architectures, such as ResNet and DeepLabV3+, are widely used for crop, field, and irrigation mapping at high spatial resolution [Safarov *et al.*, 2022; Raei *et al.*, 2022]. RNN (e.g., CNN+LSTM and 3DCNN) and attention-based (e.g., TsViT) sequence modeling has been used for crop and agricultural asset detection [Garnot *et al.*, 2020]. Transformer based models such as, ViT, Swin, and masked auto encoder (MAE) architectures improve multi-temporal land use and crop segmentation [Khan *et al.*, 2023; Chen and Li, 2025]. Recently, foundation models, such as Prithvi and geospatial embeddings like AlphaEarth, further introduce large-scale environmental contexts for agricultural analyses, including fallowing land segmentation [Brown *et al.*, 2025; Li *et al.*, 2023]. (more details in Appendix Section).

3 Methodology

3.1 Problem Formulation

We formulate agricultural land use activity mapping from remote sensing as spatio-temporal semantic segmentation. Given multi-modal observations collected over the growing season, our objective is to predict the probability of each pixel belonging to one of four land use classes: *fallow*, *pasture*, *cropland*, or *none*. For each sample, the input consists of: (i) a spatio-temporal satellite image $\mathcal{X} \in \mathbb{R}^{T\times H\times W\times C}$, where W, H, C, and T denotes satellite image width, height, channel, and selected time window; (ii) a spatial environmental context embedding from AlphaEarth $A \in \mathbb{R}^{H\times W\times D_A}$, where D_A is the dimension of the embedding space; and (iv) a land use mask $L \in \{0,1,2,3\}^{H\times W}$, indicating coarse land categories such as crop/cultivated land, developed, grassland, and forest land areas. The goal is to learn a segmentation model,

$$f_\theta : (\mathcal{X}, A, L) \rightarrow P \in [0,1]^{H\times W\times 4}. \quad (1)$$

Let $P_{i,j}$ denote the predicted class-probability vector at pixel (i,j) in the image P. Final labels are obtained via $\hat{Y}_{i,j} = \arg\max_k P^{(k)}_{i,j}$.

Given ground-truth masks $Y \in \{0,1,2,3\}^{H\times W}$, the model parameters θ are learned by minimizing loss $\mathcal{L}$ (defined in the next section):

$$\theta = \arg\min_\theta \mathcal{L}(P, Y). \quad (2)$$

3.2 Proposed Model: AG-STELLA

We propose AG-STELLA (shown in Figure 2), an agronomic knowledge-guided spatio-temporal segmentation model for agricultural land use activity mapping. The model comprises adaptive temporal window selection, environmental time-series conditioning, transformer-based spatio-temporal encoding, change-aware feature aggregation, land-aware decoding, and knowledge-guided training objectives.

Prithvi MAE Temporal Tokenizer. Given a time series image $X \in \mathbb{R}^{T\times H\times W\times C}$, we encode it using an MAE-pretrained Prithvi (300M) Vision Transformer. The encoder partitions X into patches of size (16×16) and maps them to a token sequence of length $L = T \cdot H_p \cdot W_p$, where $H_p = \frac{H}{16}, W_p = \frac{W}{16}$ with embedding dimension D, followed by stacked self-attention layers that jointly model spatial context and temporal dependencies. The output is a token representation, $Z_t \in \mathbb{R}^{L\times D}$, together with intermediate token embeddings $\{Z_t^{(\ell)}\}$ extracted from selected layers, providing hierarchical spatio-temporal representations. As the Prithvi encoder is pretrained via masked autoencoding on large-scale satellite time series with temporal positional embeddings, it yields noise-robust spatial tokens that preserve temporal ordering across monthly observations, i.e.,

$$Z_t = \phi_{\text{Pri}}(X_t), \quad Z_t \in \mathbb{R}^{L\times D},\ t = 1,\ldots,T. \tag{3}$$

Adaptive Temporal Window Selection. We adaptively select a contiguous four-month satellite image window by maximizing phenological change across the growing season. Given available months $\mathcal{M} = \{1,\ldots,T\}$, we first compute vegetation index (NDVI) maps $V_t \in \mathbb{R}^{H\times W}$ from the satellite imagery after filtering cloud-contaminated observations. We then compute temporal change, e_t, as follows:

$$e_t = \sum_{i,j}(V_{t+1}(i,j) - V_t(i,j))^2 \tag{4}$$

To select the optimal window, we maximize the temporal changes within the selected window:

$$s^* = \arg\max_{s\in\{1,\ldots,T-2\}} \sum_{r=0}^{3} e_{s+r} \tag{5}$$

This mitigates the effects of cloud contamination and missing acquisitions while preserving the contiguous temporal window that captures phenological transitions, which are essential for land use activity detection. Note that we fix the temporal window length to four months following the Prithvi encoder's fixed-length temporal clip design.

Temporal Feature Fusion (TFF). We reshape Z into a spatio-temporal token grid: $Z' \in \mathbb{R}^{T\times(H_pW_p)\times D}$. To explicitly capture phenological transitions, we compute absolute temporal differences between consecutive frames:

$$F_{temp} = \phi_{temp}(Z'_{1\ldots T}) \ \in\ \mathbb{R}^{(H_pW_p)\times D} \tag{6}$$

$$\phi_{temp}(Z'_{1\ldots T}) = \frac{1}{T-1}\sum_{t=1}^{T-1}\left|Z'_{t+1} - Z'_t\right|. \tag{7}$$

The resulting *Temporal Transition Latent Space* (F_{temp}) encodes the magnitude of temporal variation at each spatial location. This emphasizes regions undergoing strong temporal change while suppressing temporally stable background areas (shown in 3). This latent representation provides a compact and discriminative summary of land use activity dynamics.

AlphaEarth and Land Use Prior–Conditioned Decoder. We design an AlphaEarth (geospatial context-aware) and land use-aware decoder that integrates temporal dynamics, geospatial context, and coarse land use priors to produce pixel-level land use predictions. The decoder progressively upsamples the temporal transition latent map while incorporating multi-scale contextual features, AlphaEarth conditioning, and land use–based spatial regularization.

Hierarchical Spatial Context. To preserve spatial structure and boundary information, we derive context-focused skip features from intermediate encoder representations $Z^{(\ell)}$, extracted from layers $\ell \in \{2,5,8,11\}$. After temporal aggregation across the T-month horizon, the resulting skip features are reshaped into spatial maps:

$$S^{(\ell)} \in \mathbb{R}^{D\times H_p\times W_p}, \quad \ell \in \{2,5,8,11\}.$$

The change-focused latent representation (F_{temp}) is used as the decoder bottleneck, while the skip features $\{S^{(\ell)}\}$ are fused at corresponding decoding stages via feature concatenation:

$$x^{(k+1)} = \mathcal{D}\big([x^{(k)},\, S^{(\ell_k)}]\big), \tag{8}$$

where $\mathcal{D}(\cdot)$ denotes a decoder update at scale k. This skip connection facilitates gradient flow across network depths and retains high-resolution spatial cues during decoding.

AlphaEarth Conditioning (AEC). Given a decoder feature map $x^{(12)} \in \mathbb{R}^{d\times H\times W}$, we incorporate geospatial context (AlphaEarth embedding) via feature-wise linear modulation:

$$[\gamma,\beta] = \phi_{alpha}(A), \quad \gamma,\beta \in \mathbb{R}^{d\times H\times W}, \tag{9}$$

where $\phi_{alpha}(\cdot)$ denotes a learnable convolutional projection. The conditioned feature map is then computed as

$$x_{\text{alpha}} = x^{(12)} \odot \gamma + \beta, \tag{10}$$

where $\odot$ denotes elementwise multiplication. This conditioning provides spatially varying feature modulation informed by local geospatial context and ensures consistent interpretation of temporal transition signals.

Land Use–Aware Spatial Gating (LUASG). To enforce coarse spatial priors, we utilize land use mask $L^{H\times W}$ (normalized to $M \in [0,1]^{H\times W}$). Learnable soft gates modulate both contextual features and AlphaEarth conditioned feature (x_{alpha}) representations as

$$x_{\text{land+alpha}} = x_{alpha} \cdot (1 + \sigma(\phi_{land}(M))), \tag{11}$$

where $\phi_{land}(\cdot)$ denotes a channel-wise projection. This mechanism suppresses activations in non-agricultural regions while preserving smooth spatial transitions.

Segmentation Head. The final decoder features are mapped to per-pixel class logits via a lightweight prediction head composed of a 1×1 convolution, and converted to class probabilities through a softmax operation:

$$P = \text{Softmax}(\phi_{head}(x_{land+alpha})) \in [0,1]^{H\times W\times 4}, \tag{12}$$

where ϕ_{head} denotes 1×1 convolution. This formulation yields spatially consistent pixel-level predictions that integrate temporal transition cues and contextual priors.

3.3 Knowledge-Guided Training Objectives

We train the model using a composite objective that combines standard segmentation losses with two knowledge-guided regularization terms derived from hydroclimatic constraints and temporal transition dynamics:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{hyd}}\mathcal{L}_{\text{hyd}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \quad (13)$$

The pixel-wise cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=0}^{3} Y_{i,j}^{(k)} \log P_{i,j}^{(k)}, \quad (14)$$

where $P_{i,j}^{(k)}$ denotes the predicted probability of class k at pixel (i, j) and $Y_{i,j}^{(k)}$ is the ground-truth label.

Furthermore, to improve semantic segmentation under class imbalance, we apply Dice loss:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{1}{4}\sum_{k=0}^{3} \frac{2\sum_{i,j} P_{i,j}^{(k)} Y_{i,j}^{(k)} + \epsilon}{\sum_{i,j} P_{i,j}^{(k)} + \sum_{i,j} Y_{i,j}^{(k)} + \epsilon}. \quad (15)$$

Here, $P_{i,j}^{(k)}$ and $Y_{i,j}^{(k)}$ denote the predicted probability and ground-truth indicator for class k at pixel (i, j), respectively, and ϵ is a small constant added for numerical stability.

Hydroclimatic Consistency Loss. Agricultural land use decisions, particularly fallowing, are influenced by hydroclimatic conditions. To encode this prior, we construct a spatially dense ET–precipitation correlation map. Let $\{E_t(u, v)\}_{t=1}^{6}$ and $\{R_t(u, v)\}_{t=1}^{6}$ denote evapotranspiration and precipitation time series at coarse resolution. For each coarse pixel (u, v), we compute the Pearson correlation:

$$\rho(u, v) = \frac{\text{Cov}(E, R)}{\sqrt{\text{Var}(E)\,\text{Var}(R)} + \epsilon}, \quad (16)$$

where $\text{Cov}(\cdot)$ and $\text{Var}(\cdot)$ are computed over $t = 1, \ldots, 6$.

The correlation map is normalized via a sigmoid transformation and upsampled to image resolution ($\alpha \in [0, 1]^{H\times W}$). Then, using soft class probabilities $P_{i,j}^{k}$, we compute class-wise expected correlation values:

$$\mu_k = \frac{\sum_{i,j} P_{i,j}^{k}\,\alpha_{i,j}}{\sum_{i,j} P_{i,j}^{k} + \epsilon}. \quad (17)$$

We enforce the hydroclimatic constraint that fallowed regions exhibit higher ET–precipitation correlation than actively vegetated classes by minimizing:

$$\mathcal{L}_{\text{hyd}} = \sum_{k\in\{\text{crop,pasture}\}} \max\big(0, m - (\mu_{\text{fallow}} - \mu_k)\big), \quad (18)$$

where m is a margin. We set $m = 0.2$, based on empirical separation observed in the training set between class-conditional hydroclimatic statistics.

Temporal Transition Latent Space Loss. To guide the Temporal Transition Latent Space to encode meaningful seasonal dynamics, we provide auxiliary supervision derived from vegetation index variations. Given NDVI maps $\{V_t\}_{t=1}^{T}$, we compute the average temporal transition magnitude:

$$\Delta V = \phi_{vi}(V_{1\ldots T}) \ \in\ \mathbb{R}^{(H\times W)} \quad (19)$$

$$\phi_{vi}(V_{1\ldots T}) = \frac{1}{T-1}\sum_{t=1}^{T-1} |V_{t+1} - V_t|. \quad (20)$$

The transition map ΔV is downsampled to $\Delta V' \in \mathbb{R}^{(H_p W_p)}$ to match the spatial resolution of F_{temp}. We then minimize the following consistency loss:

$$\mathcal{L}_{\text{temp}} = \|\text{Norm}(F_{\text{temp}}) - \text{Norm}(\Delta V')\|_2^2\,, \quad (21)$$

where $\text{Norm}(\cdot)$ denotes per-sample normalization.

4 Experimentation and Results

4.1 Dataset Details

We construct a multi-source agricultural land use activity dataset across three U.S. states: Utah (UT), Colorado (CO), and Washington (WA). land use labels are obtained from state-specific authoritative sources, including the Utah Water-Related Land Use (WRLU) dataset (2023)[1], the Washington State Department of Agriculture Agricultural Land Use dataset[2], and the Colorado Division of Water Resources GIS dataset[3]. The study period spans 2018–2020 for WA, 2020 for CO, and 2024 for UT. Original land use categories are unified into three activity classes (fallow, pasture, and cropland) for consistency across states.

We collect monthly Sentinel-2 surface reflectance imagery (April to September) via Google Earth Engine [4] for all study years (applying cloud filtering of 50% or lower). Following standard remote sensing procedures, images are segmented into non-overlapping patches of size 224×224 pixels [Hoque *et al.*, 2025]. We retain only patches overlapping labeled agricultural regions, using land-cover masks derived from state data and the National Land Cover Database [5]. Auxiliary climate variables, including evapotranspiration (ET; 1 km resolution)[6] and precipitation (800 m resolution)[7], are spatially aligned to satellite pixels using nearest-neighbor matching. All satellite imagery, labels, and auxiliary variables are co-registered to generate aligned multi-modal training samples. The constructed dataset contains 3129 samples for CO, 4366 for UT, and 13861 for WA, with strong class imbalances; for example, UT and CO contain only 6.9% and 9.7% fallowed land (detailed in Appendix).

[1] https://dwre-utahdnr.opendata.arcgis.com/pages/wrlu-data
[2] https://agr.wa.gov/departments/land-and-water/natural-resources/agricultural-landuse
[3] https://dwr.colorado.gov/services/data-information/gis
[4] https://earthengine.google.com/
[5] https://www.mrlc.gov/
[6] https://earlywarning.usgs.gov/fews/product/460/
[7] https://prism.oregonstate.edu/

Model	UT				CO				WA			
	mAcc(%)	F1(%)			mAcc(%)	F1(%)			mAcc(%)	F1(%)		
		Fallow	Pasture	Crop		Fallow	Pasture	Crop		Fallow	Pasture	Crop
ResNet	51.7	31.2	27.0	68.0	59.8	35.4	49.1	67.4	68.7	68.8	48.5	74.5
ViT	55.5	34.1	36.0	70.2	58.5	38.6	45.2	66.1	65.1	67.5	34.3	74.4
Swin	59.4	37.0	39.3	75.6	64.0	46.0	52.2	69.4	69.9	72.7	41.5	80.3
DeepLabV3+	53.8	28.8	33.5	71.9	60.5	37.1	50.0	69.3	62.3	63.3	27.5	72.9
CLIP	48.4	25.9	31.5	59.0	47.0	29.0	32.6	49.1	57.4	59.1	20.1	67.0
TSViT	47.7	29.2	22.4	68.1	68.1	14.1	50.7	64.5	55.9	53.5	8.3	70.5
VideoMAE	55.7	28.9	37.5	72.6	62.6	37.6	53.6	69.2	62.6	57.4	36.6	70.2
UNet3D	64.3	**47.2**	47.1	**80.2**	66.7	41.7	59.8	**75.6**	62.6	59.3	32.3	74.5
ConvLSTM	56.2	33.7	37.0	76.4	61.4	32.5	54.7	74.2	58.4	52.7	28.4	71.2
Prithvi	63.1	42.7	46.9	77.2	67.0	42.4	62.9	75.1	74.1	74.3	58.6	79.4
AlphaEarth+MLP	61.4	41.9	43.2	79.5	66.6	37.3	65.5	79.1	71.1	72.2	46.8	80.1
AlphaEarth+CNN	62.6	45.4	43.8	79.6	68.9	42.0	66.8	78.8	73.1	75.2	48.5	81.6
AG-STELLA	**70.0**	46.8	**50.5**	78.0	**74.3**	**49.1**	**67.8**	74.5	**85.4**	**84.7**	**74.4**	**87.4**

Table 1: Micro accuracy and class-wise F1-scores across three states under state-specific training. Best (dark green) and second-best (light green) results are highlighted using color coding.

TFF	AEC	LUASG	HCL	TTLSP	F1 (Macro)
×	×	×	×	×	55.1
×	✓	✓	✓	✓	63.8
✓	×	✓	✓	✓	61.5
✓	✓	×	✓	✓	64.5
✓	✓	✓	×	✓	64.1
✓	✓	✓	✓	×	59.3
✓	✓	✓	✓	✓	**66.0**

Table 2: Ablation study of AG-STELLA components. We evaluate the contribution of five key components: (i) Temporal Feature Fusion (TFF), (ii) AlphaEarth Conditioning (AEC), (iii) Land Use–Aware Spatial Gating (LUASG), (iv) Hydroclimatic Consistency Loss (HCL), and (v) Temporal Transition Latent Space Loss (TTLSP), by removing one component at a time. Performance is reported using macro-averaged F1-score (%).

State	Model	mAcc.	F1-score		
			Fallow	Pasture	Crop
UT	Baseline	70.0	47.2	50.5	80.2
	Prithvi	61.8	41.4	**46.2**	76.5
	AlphaEarth+MLP	60.0	39.8	43.0	79.2
	AlphaEarth+CNN	59.7	37.3	42.8	**79.3**
	AG-STELLA	**63.5**	**43.4**	45.0	77.6
CO	Baseline	74.3	49.1	67.8	75.6
	Prithvi	67.1	40.5	63.8	75.0
	AlphaEarth+MLP	64.0	32.0	62.4	78.2
	AlphaEarth+CNN	67.0	36.6	64.3	**78.4**
	AG-STELLA	**70.3**	**44.4**	**65.1**	74.3
WA	Baseline	85.4	84.7	74.4	87.4
	Prithvi	75.2	75.2	59.8	80.1
	AlphaEarth+MLP	69.0	70.9	42.3	79.7
	AlphaEarth+CNN	71.9	74.0	47.6	82.3
	AG-STELLA	**84.6**	**82.3**	**72.6**	**85.7**

Table 3: Cross-state performance comparison across models. 'Baseline' indicates the state-wise best performance. Macro accuracy (mAcc.) and class-wise F1-scores are reported for Fallow, Pasture, and Crop.

4.2 Experimentation Setting

We split each state's dataset into 70% for training, 15% for validation, and 15% for testing. We follow a **spatially disjoint** [Mandal *et al.*, 2025] partitioning strategy to avoid spatial leakage and to ensure that training, validation, and test samples originate from non-overlapping geographic regions.

Model implementation was conducted using PyTorch and executed on an NVIDIA A40 GPU. We initialize the spatio-temporal encoder with a pretrained Prithvi MAE model and fine-tune only the last six transformer layers, while keeping earlier layers frozen to balance adaptation and overfitting.

We optimize hyperparameters via grid search over learning rates $\{10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$, batch sizes $\{8, 16\}$, and loss weighting coefficients $\lambda_{\text{Dice}}, \lambda_{\text{hyd}}, \lambda_{\text{temp}} \in \{0.1, 0.2, 0.5, 0.7\}$. The optimal configuration is selected based on macro-F1 performance on the validation set (More details in Appendix).

4.3 Result Analysis

Comparison with Baseline Models. We employ state-specific training and evaluate the models independently for each state to capture regional agricultural heterogeneity. Our model is compared against 12 fully fine-tuned baselines, including vision-only models (ResNet, ViT, Swin, DeepLabV3+, CLIP), temporal vision models (TSViT, VideoMAE, UNet3D, ConvLSTM), and Earth-observation foundation models (Prithvi, AlphaEarth+MLP, AlphaEarth+CNN). All baselines are trained on identical splits using the same supervision and evaluation protocol to ensure a controlled comparison.

Across all states, our model achieves the highest **macro accuracy (mAcc)** and class-wise F1-score (shown in Table 1), with the most pronounced improvements observed for **fallow** and **pasture** (the two most challenging and underrepresented agricultural land use activity classes). Compared to the strongest vision and temporal baselines, our approach improves macro accuracy from **64.3% to 70.0% in Utah** (+8.9% relative), from **68.9% to 74.3% in Colorado** (+7.8%), and from **74.1% to 85.4% in Washington** (+15.3%). These consistent improvements in minority-class performance (including substantial increases in Fallow and Pasture F1-score) indicates improved discrimination of

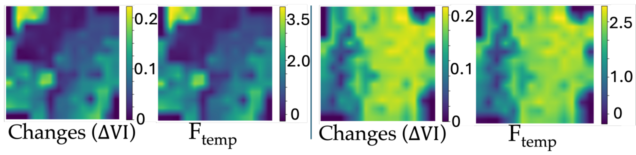

Figure 3: Shows two examples where the temporal changes are captured by model (F_{temp}) and observation from vegetation indices (ΔVI). In both examples, correlations are higher than 0.95.

land use activity beyond static appearance or coarse temporal trends (shown in Figure 3). Note that performance for fallow and pasture remains comparatively lower in Utah and Colorado due to strong spectral similarity with active cropland and low prevalence (severe class imbalance in the dataset). We further analyze performance against foundation models. Relative to Prithvi, our method improves macro accuracy by **+11.0% (UT)**, **+10.9% (CO)**, and **+15.3% (WA)**, with notable gains in Fallow and Pasture F1-score across all regions. Compared to **AlphaEarth**-based models, our approach achieves additional improvements of **8–16% in macro accuracy**, with especially large gains for pasture, including an increase from **58.6% to 74.4% (+27.0%)** in Washington. These results suggest that while foundation models provide strong generic representations, they remain limited in resolving fallow–pasture ambiguity. By explicitly modeling land use activity, our approach delivers more balanced and region-robust performance, particularly for fallow and pasture (more details in Appendix).

Transfer Learning for Cross-State Agricultural Land Use Activity Mapping. In the cross-state setting, models are trained on data pooled from all three states. Prior studies have shown that strong state-level heterogeneity in agricultural practices often favors state-specific training for optimal performance [Mandal *et al.*, 2025]. We therefore evaluate whether cross-state training can transfer effectively and achieve performance comparable to state-specific models. In Table 3, the baseline denotes the best state-specific model and serves as an approximate upper bound. Compared to its own state-specific performance (Table 1), AG-STELLA exhibits a markedly smaller fallow performance drop (-3.4% UT, -4.7% CO, -2.4% WA) compared to other models. In particular, models based on Prithvi and AlphaEarth suffer fallow degradation (e.g. Prithvi: 41.4% UT, 40.5% CO, 75.2% WA; AlphaEarth+CNN: 37.3% UT, 36.6% CO, 74.0% WA). In contrast, AG-STELLA achieves F1-score of **43.4% (UT), 44.4% (CO), and 82.3% (WA)** for fallow land (outperforming the comparing models by **+4.8%, +9.6%, and +9.6%**, respectively). This demonstrates AG-STELLA's ability to transfer knowledge across the states.

Error Analysis (Confusing Crop Types). To characterize dominant failure modes, we analyze crop types most frequently misclassified as Fallow or Pasture. For each state, we identify the crop types along with the proportion of pixels predicted as these classes (See Appendix). Across states, wheat-related crops (e.g., winter wheat and spring grain) consistently appear among the top fallow confusions, particularly in Colorado and Washington (potentially reflecting similar spectral and temporal signatures during dormant or post-harvest periods). In contrast, grass and hay dominated crops (e.g., bluegrass and grass hay) dominate the top pasture misclassifications (likely because their sustained herbaceous cover, limited intra-seasonal spectral variation, and management practices closely resemble actively grazed pasture systems)(see details in Appendix).

5 Ablation Studies

We conduct an ablation study to quantify the contribution of each architectural and loss component in AG-STELLA. Starting from the full model, we systemically remove one component at a time while keeping all other settings fixed. As shown in Table 2, each component contributes positively to performance, with the full model achieving the highest macro F1-score. In particular, removing TTLSP or TFF leads to the largest performance degradation which highlights the importance of explicitly modeling temporal land use transitions and multi-temporal feature fusion. Furthermore, ablating AEC, LUASG, or HCL consistently reduces performance, indicating that AlphaEarth conditioning, land use–aware spatial gating, and hydroclimatic consistency jointly improve robustness and generalization.

6 Conclusion

In this work, we present AG-STELLA, a knowledge-guided, end-to-end trainable spatio-temporal model for agricultural land use activity mapping. Our model integrates temporal vegetation dynamics, AlphaEarth geospatial context, and hydroclimatic constraints. Experiments across three U.S. states show consistent gains over strong spatio-temporal baselines and demonstrates the effectiveness of combining pre-trained foundation models with domain-informed inductive biases. Beyond accuracy, AG-STELLA aligns predictions with hydro-climatic constraints which enables scalable agricultural monitoring in water-stressed regions. Moreover, AG-STELLA supports practical applications, including drought response, water allocation, and groundwater management through monitoring seasonal land-use transitions. It can also aid agricultural policy verification and analysis of fallowing behavior impacting regional crop production and economic planning.

Despite these promising results, we assume a single dominant agricultural land use activity within a growing season, which does not capture cross-season or within-year transitions such as mid-season fallowing. Pasture land remains inherently ambiguous, as it can be managed under both irrigated and non-irrigated regimes that lead to overlapping temporal and hydro-climatic signatures. Additionally, hydro-climatic constraints are derived from coarse-resolution ET and precipitation data, which may not fully reflect field-scale management variability. Therefore, future work will focus on cross-season and multi-year modeling to capture land use transitions across planting cycles. Incorporating irrigation intensity signals, higher-resolution hydro-climatic data, and probabilistic or multi-label formulations, along with scaling via weak supervision and active learning, offers a promising path toward large-scale operational agricultural monitoring.

Acknowledgments

This material is based upon work supported by the AI Research Institutes program supported by NSF and USDA-NIFA under the AI Institute: Agricultural AI for Transforming Workforce and Decision Support (AgAID) award No. 2021-67021-35344. This work was partially supported by University of Virginia Strategic Investment Fund award number SIF160.

References

[Bak *et al.*, 2025] Hyeok-Jin Bak, Eun-Ji Kim, Ji-Hyeon Lee, Sungyul Chang, Dongwon Kwon, Woo-Jin Im, Woon-Ha Hwang, Jae-Ki Chang, Nam-Jin Chung, and Wan-Gyu Sang. Deep learning-based semantic segmentation for rice yield estimation by analyzing the dynamic change of panicle coverage. *Frontiers in Plant Science*, 16:1611653, 2025.

[Brown *et al.*, 2025] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

[Chen and Li, 2025] Miaomiao Chen and Lianfa Li. Hierarchical transfer learning with transformers to improve semantic segmentation in remote sensing land use. *Remote Sensing*, 17(2):290, 2025.

[Chen *et al.*, 2023] Hong Chen, Sha Chen, Runjia Yang, Liping Shan, Jinmin Hao, and Yanmei Ye. Optimizing the cropland fallow for water resource security in the groundwater funnel area of china. *Land*, 12(2):462, 2023.

[Cheng *et al.*, 2021] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875, 2021.

[Christofi *et al.*, 2025] Konstantinos Christofi, Charalambos Chrysostomou, Iason Tsardanidis, Michalis Mavrovouniotis, Giorgia Guerrisi, Charalampos Kontoes, and Diofantos G Hadjimitsis. Remote sensing of grasslands: Performance comparison of radar and optical data in machine learning classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:295–300, 2025.

[Di Cicco *et al.*, 2017] Maurilio Di Cicco, Ciro Potena, Giorgio Grisetti, and Alberto Pretto. Automatic model based dataset generation for fast and accurate crop and weeds detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5188–5195. IEEE, 2017.

[Eslamifar *et al.*, 2024] Gholamreza Eslamifar, Hamid Balali, and Alexander Fernald. Fallowing strategy and its impact on surface water and groundwater withdrawal, and agricultural economics: A system dynamics approach in southern new mexico. *Water*, 16(1):181, 2024.

[Garnot *et al.*, 2020] Vivien Sainte Fare Garnot, Loic Landrieu, Sebastien Giordano, and Nesrine Chehata. Satellite image time series classification with pixel-set encoders and temporal self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12325–12334, 2020.

[Greiner *et al.*, 2021] Clemens Greiner, Hauke-Peter Vehrs, and Michael Bollig. Land-use and land-cover changes in pastoral drylands: Long-term dynamics, economic change, and shifting socioecological frontiers in baringo, kenya. *Human Ecology*, 49(5):565–577, 2021.

[Hoque *et al.*, 2025] Oishee Bintey Hoque, Nibir Chandra Mandal, Abhijin Adiga, Samarth Swarup, Sayjro Kossi Nouwakpo, Amanda Wilson, and Madhav Marathe. Knowledge-informed deep learning for irrigation type mapping from remote sensing. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9692–9700. International Joint Conferences on Artificial Intelligence Organization, 8 2025. AI and Social Good.

[Hung *et al.*, 2014] Calvin Hung, Zhe Xu, and Salah Sukkarieh. Feature learning based approach for weed classification using high resolution aerial images from a digital camera mounted on a uav. *Remote Sensing*, 6(12):12037–12054, 2014.

[Khan *et al.*, 2023] Asim Hameed Khan, Zuhair Zafar, Muhammad Shahzad, Karsten Berns, and Muhammad Moazam Fraz. Crop type classification using multitemporal sentinel-2 satellite imagery: A deep semantic segmentation approach. In *2023 International Conference on Robotics and Automation in Industry (ICRAI)*, pages 1–6. IEEE, 2023.

[Koley and Chockalingam, 2025] Swadhina Koley and Jeganathan Chockalingam. Extraction of fallow lands in the rabi season using time-series landsat 8 imageries. In *Revealing Ecosystem Services Through Geospatial Technologies: Beyond the Surface*, pages 41–57. Springer, 2025.

[Lei *et al.*, 2024] Lian Lei, Qiliang Yang, Ling Yang, Tao Shen, Ruoxi Wang, and Chengbiao Fu. Deep learning implementation of image segmentation in agricultural applications: a comprehensive review. *Artificial Intelligence Review*, 57(6):149, 2024.

[Li *et al.*, 2023] Hanxi (Steve) Li, Sam Khallaghi, Michael Cecil, Fatemeh Kordi, Paolo Fraccaro, Hamed Alemohammad, and Rahul Ramachandran. HLS Multi Temporal Crop Classification Model, August 2023.

[Liu *et al.*, 2025] Tao Liu, Le Yu, Xiaoxuan Liu, Dailiang Peng, Xin Chen, Zhenrong Du, Ying Tu, Hui Wu, and Qiang Zhao. A global review of monitoring cropland abandonment using remote sensing: Temporal–spatial patterns, causes, ecological effects, and future prospects. *Journal of Remote Sensing*, 5:0584, 2025.

[Mandal *et al.*, 2025] Nibir Chandra Mandal, Oishee Bintey Hoque, Mandy L Wilson, Samarth Swarup, Sayjro K Nouwakpo, Abhijin Adiga, and Madhav Marathe. Irrisight: A large-scale multimodal dataset and scalable pipeline to address irrigation and water management in agriculture. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[Mirzaei and Garcia, 2024] Shokoufeh Mirzaei and Anna Maria Garcia. A framework to define optimal fallowing programs: a case study of california water system. 2024.

[Morais *et al.*, 2024] Tiago G Morais, Tiago Domingos, João Falcão, Manuel Camacho, Ana Marques, Inês Neves, Hugo Lopes, and Ricardo FM Teixeira. Permanent pastures identification in portugal using remote sensing and multi-level machine learning. *Frontiers in Remote Sensing*, 5:1459000, 2024.

[Mortensen *et al.*, 2016] Anders Krogh Mortensen, Mads Dyrmann, Henrik Karstoft, Rasmus Nyholm Jørgensen, and René Gislum. Semantic segmentation of mixed crops using deep convolutional neural network. 2016.

[Noorazar *et al.*, 2025] Hossein Noorazar, Michael P Brady, Supriya Savalkar, Amin Norouzi Kandelati, Mingliang Liu, Perry Beale, Andrew M McGuire, Timothy Waters, and Kirti Rajagopalan. Monitoring double-cropped extent with remote sensing in areas with high crop diversity. *Plants*, 14(9):1362, 2025.

[Norton *et al.*, 2021] Cynthia L Norton, Matthew P Dannenberg, Dong Yan, Cynthia SA Wallace, Jesus R Rodriguez, Seth M Munson, Willem JD van Leeuwen, and William K Smith. Climate and socioeconomic factors drive irrigated agriculture dynamics in the lower colorado river basin. *Remote Sensing*, 13(9):1659, 2021.

[Paolini *et al.*, 2022] Giovanni Paolini, Maria Jose Escorihuela, Olivier Merlin, Magí Pamies Sans, and Joaquim Bellvert. Classification of different irrigation systems at field scale using time-series of remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:10055–10072, 2022.

[Raei *et al.*, 2022] Ehsan Raei, Ata Akbari Asanjan, Mohammad Reza Nikoo, Mojtaba Sadegh, Shokoufeh Pourshahabi, and Jan Franklin Adamowski. A deep learning image segmentation model for agricultural irrigation system classification. *Computers and Electronics in Agriculture*, 198:106977, 2022.

[Rosanna *et al.*, 2025] Paolino Rosanna, Sabia Emilio, Braghieri Ada, Riviezzi Amelia Maria, Vignozzi Luca, Claps Salvatore, Baldassarre Daniele, Pacelli Corrado, Caparra Pasquale, and Di Trana Adriana. Methodological approaches for estimating the biomass of natural pastures in the lucanian hills using uav remote sensing. In *International Workshop IFToMM for Sustainable Development Goals*, pages 641–648. Springer, 2025.

[Safarov *et al.*, 2022] Furkat Safarov, Kuchkorov Temurbek, Djumanov Jamoljon, Ochilov Temur, Jean Chamberlain Chedjou, Akmalbek Bobomirzaevich Abdusalomov, and Young-Im Cho. Improved agricultural field segmentation in satellite imagery using tl-resunet architecture. *Sensors*, 22(24):9784, 2022.

[Satapathy *et al.*, 2024] Trupti Satapathy, Jörg Dietrich, and Meenu Ramadas. Agricultural drought monitoring and early warning at the regional scale using a remote sensing-based combined index. *Environmental Monitoring and Assessment*, 196(11):1132, 2024.

[Song *et al.*, 2022] Wen Song, Alexander V Prishchepov, and Wei Song. Mapping the spatial and temporal patterns of fallow land in mountainous regions of china. *International Journal of Digital Earth*, 15(1):2148–2167, 2022.

[Tortajada *et al.*, 2017] Cecilia Tortajada, Matthew J Kastner, Joost Buurman, and Asit K Biswas. The california drought: Coping responses and resilience building. *Environmental Science & Policy*, 78:97–113, 2017.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[WA State Legislature, 2018] WA State Legislature. Washington administrative code § 173-166-040: Emergency drought relief. https://app.leg.wa.gov/WAC/default.aspx?cite=173-166, 2018. Washington Administrative Code.

[Wang, 2023] Haoying Wang. Going back to grassland? assessing the impact of groundwater decline on irrigated agriculture using remote sensing data. *Remote Sensing*, 15(6):1698, 2023.

[Weiss *et al.*, 2020] Marie Weiss, Frédéric Jacob, and Grgory Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, 236:111402, 2020.

[WWAP, 2019] WWAP. Leaving no one behind. https://unesdoc.unesco.org/ark:/48223/pf0000367306, 2019. United Nations World Water Assessment Programme, Technical Report.

[Żarczyński *et al.*, 2023] Piotr Jarosław Żarczyński, Sławomir Józef Krzebietke, Stanisław Sienkiewicz, and Jadwiga Wierzbowska. The role of fallows in sustainable development. *Agriculture*, 13(12):2174, 2023.