SPEAR-SIM: PROCESS-GUIDED WEAK SUPERVISION FOR AGRICULTURAL FOUNDATION MODELS USING MECHANISTIC CROP SIMULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth observation foundation models have demonstrated strong transferability across remote sensing tasks, yet adapting them to agricultural applications remains challenging because large-scale supervision for intermediate crop physiological processes is largely unavailable. Existing downstream adaptation therefore relies primarily on sparse labels such as crop yield, providing limited guidance for learning physiologically meaningful representations. An observation-constrained process-guided adaptation framework, SPEAR-SIM, is proposed to leverage mechanistic crop simulations as weak supervision for Earth observation foundation models. The SPEAR-SIM framework first generates ensembles of APSIM crop growth trajectories and selects simulations that best agree with satellite-observed crop development, producing observationally validated teacher signals for continued self-supervised adaptation of a pretrained Earth observation foundation model. The resulting representations are evaluated through downstream U.S. county-level crop yield prediction across multiple forecasting cutoffs. Experimental results demonstrate consistent improvements over both handcrafted-feature baselines and conventional foundation model adaptation. The proposed framework increases early-season corn prediction performance from an R^2 of 0.308 to 0.623, evaluated across states that include regions without simulator-derived supervision, while improving soybean prediction at five of seven forecasting cutoffs, most consistently from mid-summer onward. Additional analyses further show that the relative value of different physiological auxiliary targets varies by crop and by stage of the growing season, with no single subset of targets dominating throughout. More broadly, these findings demonstrate that observation-constrained process-guided weak supervision provides an effective paradigm for integrating mechanistic crop simulations with foundation models and offers a general strategy for combining scientific simulators with modern self-supervised learning.

1 INTRODUCTION

Earth observation (EO) foundation models (Cong et al., 2022; Tseng et al., 2023; Feng et al., 2026) have recently demonstrated remarkable transferability across a wide range of remote sensing tasks by learning generic representations from large-scale unlabeled satellite imagery. While these pretrained models substantially reduce the dependence on task-specific training data, adapting them to agricultural applications remains challenging. Most downstream adaptation relies on sparse supervisory signals such as crop type or end-of-season yield, providing little guidance about the underlying physiological processes that govern crop growth. As a result, learned representations often capture correlations with yield rather than the temporal dynamics of crop development that determine it.

Learning physiologically meaningful crop representations ideally requires supervision from intermediate crop variables such as leaf area index (LAI), biomass accumulation, phenological development, and plant water stress. These variables directly describe crop growth throughout the season and therefore provide substantially richer supervisory signals than a single end-of-season yield observation. However, large-scale field measurements of these physiological variables are rarely avail-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

able across diverse geographic regions and growing seasons, making direct supervision impractical for foundation model adaptation.

Mechanistic crop growth models such as APSIM (Holzworth et al., 2018) simulate crop development by integrating weather, soil, genotype, and management information, providing physically interpretable trajectories of crop growth variables throughout the growing season. Although these simulators have been widely used for agricultural forecasting and management studies, they are rarely incorporated into the representation learning process of modern EO foundation models. Furthermore, simulation outputs may deviate from real-world crop development because of uncertainty in environmental conditions and management practices, limiting their usefulness as direct supervisory signals without additional validation.

In this work, we propose an observation-constrained process-guided adaptation framework for agricultural foundation models. Rather than directly supervising the model with simulated trajectories, we generate ensembles of APSIM crop growth simulations and identify those that best agree with satellite-observed crop development. These trajectories then serve as weak supervision for a training-time-only adaptation stage: we continue self-supervised pretraining of a pretrained EO foundation model's temporal encoder while a small auxiliary process head, added only for this stage and discarded afterward, regresses the selected teacher trajectories; no yield label is used anywhere in this stage. County-level yield prediction is evaluated separately by a lightweight probe on the frozen adapted representation, isolating the contribution of simulator supervision and, since physiological ground truth is otherwise unavailable, serving as our indirect measure of representation quality.

We build on the SPEAR-NeXT backbone (Ranjan et al., 2026b), a causal multi-horizon latent forecaster over per-timestep SPEAR states (Ranjan et al., 2026a), and on the AgroBench county-yield benchmark (Singh et al., 2026), evaluating our adapted representation against the unmodified backbone under an identical protocol.

The primary contributions of this work are as follows:

- We propose an observation-constrained process-guided adaptation framework that turns mechanistic crop simulations into training-time-only weak supervision for a pretrained agricultural Earth observation foundation model, with an explicit separation between label-free simulator-guided adaptation and yield probing.
- We introduce an observation-constrained teacher selection strategy that aligns simulated crop trajectories with satellite observations before representation learning, with statistics and a sensitivity analysis over 23,592 APSIM simulations spanning 934 corn and 929 soybean county-seasons.
- We demonstrate that process-guided adaptation improves downstream yield prediction over the same backbone at every corn cutoff and five of seven soybean cutoffs, with the largest gains early in the corn season and extending to states without simulator labels; the most useful physiological targets vary by crop and growth stage.

2 Related work

EO foundation models. Masked reconstruction, contrastive alignment and temporal summarization dominate EO pretraining (Cong et al., 2022; Fuller et al., 2023; Tseng et al., 2023; Feng et al., 2026; Brown et al., 2025). SPEAR learns wavelength-conditioned per-pixel spectral states for optical, radar and climate inputs (Ranjan et al., 2026a); SPEAR-NeXT adds a causal temporal transformer trained to forecast future SPEAR states at multiple horizons (Ranjan et al., 2026b). These models are adapted downstream with task labels only; we study adding simulator knowledge during adaptation.

Knowledge-guided ML, benchmarks, and privileged supervision. Theory-guided data science injects scientific models into learning through loss terms, architectures, or simulated pretraining data (Karpatne et al., 2017; Jia et al., 2019). In agriculture, KGML-ag couples process-model structure and simulations with neural networks (Liu et al., 2022), crop-model outputs have been used as features for yield regressors (Shahhosseini et al., 2021), and crop simulations have supervised intermediate variables of yield models (He et al., 2023); physics-guided pretraining has also been applied

to foundation models in aquatic science (Yu et al., 2025). SPEAR-SIM differs in three respects: it filters simulations by satellite observations, adapts a *pretrained EO foundation model* rather than a task-specific network, and uses simulator output only during a *label-free adaptation stage*, never at inference. County-level yield prediction from remote sensing has a long history (You et al., 2017), with benchmarks such as SustainBench, CropNet and CY-Bench (Yeh et al., 2021; Lin et al., 2024; Kallenberg et al., 2026) and multimodal transformers such as MMST-ViT (Lin et al., 2023); we use AgroBench (Singh et al., 2026) for its pixel-level multimodal sequences matching the backbone's input. Multitask learning regularizes shared representations with related targets (Caruana, 1997), and learning using privileged information exploits training-time-only signals (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016); simulator trajectories are privileged in this sense.

3 Background: backbone and benchmark

SPEAR-NeXT backbone (Ranjan et al., 2026b). For each modality $m \in \{\mathrm{S2}, \mathrm{S1}, \mathrm{clim}\}$ (Sentinel-2 optical, Sentinel-1 radar, and climate/terrain inputs), a frozen SPEAR encoder maps the observation at timestep t to a D-dimensional spectral state $\boldsymbol{e}_t^m \in \mathbb{R}^D$, with $D = 32$ (Ranjan et al., 2026a). A causal transformer q_θ^m (4 layers, width 128, RoPE, QK-norm, SwiGLU, month/year embeddings) produces contextual states $\boldsymbol{r}_t^m = q_\theta^m(\boldsymbol{e}_{1:t}^m)$ that depend only on observations up to t, and a set of K=46 horizon-specific prediction heads g_k^m forecast future states $\hat{\boldsymbol{e}}_{t+k|t}^m = g_k^m(\boldsymbol{r}_t^m)$ at horizon $k \in \{1, \ldots, K\}$ timesteps ahead, trained with a horizon-weighted objective $\mathcal{L}_{\mathrm{NEP}}$ that combines a directional (cosine) term with a magnitude (squared-error) term and down-weights longer horizons, averaged over valid context–target pairs with stop-gradient targets (restated in full in Appendix G for reference). We use the released U.S. checkpoints for all three modalities; the backbone architecture and pretraining are described in Ranjan et al. (2026b) and are not contributions of this paper.

Yield feature. Following the SPEAR-NeXT yield protocol, pixel sequences are encoded and averaged to county-season level. Each county-season sequence has $T = 15$ timesteps (fourteen-day composites, April through October; see below). At forecast cutoff $t_c \in \{1, \ldots, T\}$ the county feature concatenates the mean contextual state over $t \le t_c$ for the three modalities (3×128), each modality's horizon-head forecast of the end-of-season state at horizon $k = \max(1, T - t_c)$ (3×32), and five NDVI/EVI summaries up to t_c, giving $\boldsymbol{x}_{t_c} \in \mathbb{R}^{485}$. Because q_θ is causal, $\boldsymbol{x}_{t_c}$ uses no imagery from after the cutoff.

AgroBench (Singh et al., 2026). AgroBench links USDA National Agricultural Statistics Service county yields to Cropland-Data-Layer-masked pixels (up to 300 pixels per county-season) with $T = 15$ fourteen-day composites from April to October of Sentinel-2 (10 bands), Sentinel-1 (VV/VH polarizations), MODIS land-surface temperature, CHIRPS precipitation and SRTM elevation, spanning 2017–2024. We use corn, evaluated across eight states (Iowa, Illinois, Indiana, Michigan, Minnesota, Missouri, Ohio and Wisconsin), and soybean, evaluated across eleven states (Iowa, Illinois, Indiana, Kansas, Louisiana, Michigan, Minnesota, Missouri, Mississippi, Nebraska and Wisconsin). County yield is a county-level target: we never treat it as a pixel-level label.

4 Method

4.1 Framework Overview

SPEAR-SIM adapts SPEAR-NeXT (Ranjan et al., 2026b) using mechanistic crop simulations as training-time supervision: rather than treating simulator outputs as ground truth, it generates diverse crop growth trajectories, retains those consistent with satellite observations, and uses the resulting teacher labels to guide continued pretraining of the foundation model's temporal encoder. It follows the three stages summarized above: teacher generation (Section 4.2), observation-constrained selection (Section 4.3), and process-guided adaptation followed by separate yield probing (Sections 4.4–4.5).

Figure 1 provides an overview of the complete SPEAR-SIM framework, while the following subsections describe each stage in detail.

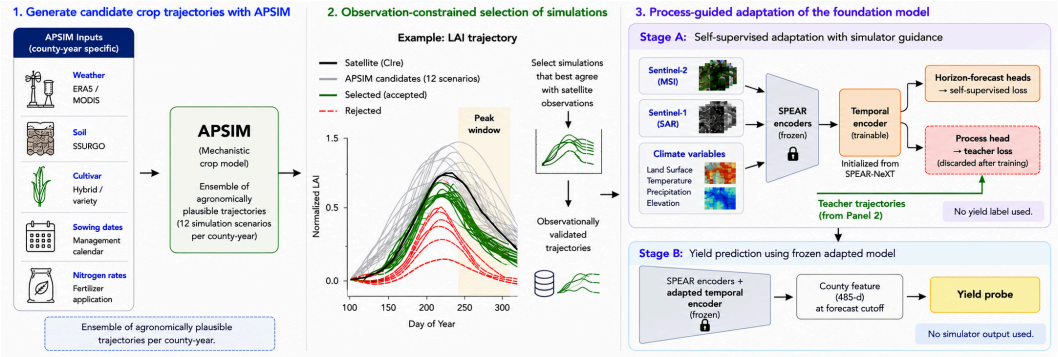


Figure 1: Overview of the SPEAR-SIM framework. (1) County-specific APSIM simulations generate an ensemble of candidate physiological trajectories under diverse management scenarios. (2) Observation-constrained filtering retains only the trajectories consistent with satellite-observed canopy development, producing consensus teacher labels. (3) The teacher labels supervise an auxiliary process head during continued self-supervised pretraining of the SPEAR-NeXT temporal encoder; the process head is discarded after this stage, and no yield label is used at this or any earlier stage. Yield prediction (Section 4.5) is trained separately on the resulting frozen representation and uses no simulator output at inference.

4.2 Agronomic Teacher Generation

The first stage of the proposed framework constructs process-level supervisory signals using APSIM Next Generation (Holzworth et al., 2018), a widely adopted mechanistic crop growth model that simulates crop development by integrating weather, soil properties, crop genetics, and management practices. APSIM has been extensively validated across diverse cropping systems and produces physically consistent trajectories of key crop growth variables throughout the growing season, making it well suited for generating candidate supervisory signals for agricultural representation learning (Winn et al., 2023).

Rather than relying on a single deterministic simulation, we generate an ensemble of agronomically plausible crop growth trajectories for every county-season by varying cultivar selection, planting date, and nitrogen management while conditioning on county-specific weather and soil information. The resulting ensemble captures uncertainty arising from management practices while preserving the biological consistency enforced by the underlying crop growth model. Weather forcing comes from NASA POWER (Sparks, 2018); soil comes from APSIM template profiles for corn and from SSURGO (USDA Natural Resources Conservation Service, 2024) for soybean (Appendix A discusses this crop-specific choice).

The simulator inputs are configured using geographically appropriate cultivar assignments and management practices, assigning counties to one of nine latitude zones spanning Iowa, Illinois, and Indiana (Appendix A, Figure 4). Each county-season is simulated under an ensemble $\mathcal{T}_{c,y} = \{T_i\}$ of twelve management scenarios, spanning four zone-specific sowing dates and three nitrogen application rates. This yields 23,592 simulations covering 934 corn and 929 soybean county-seasons.

Each simulated trajectory provides process variables including LAI, above-ground biomass, grain weight, water stress, and phenological stage, together forming a candidate teacher bank later refined using satellite observations.

4.3 Observation-Constrained Teacher Selection

Uncertainty in management practices can produce multiple plausible simulations for the same county-season; rather than treating every simulation equally, we identify those that best agree with satellite-observed crop development before using them for representation learning.

We compute the canopy proxy from the Sentinel-2 red-edge chlorophyll index, $\mathrm{CI}_{\mathrm{re}} = B7/B5 - 1$, which has been shown to track canopy development and leaf area dynamics in agricultural systems (Clevers & Gitelson, 2013). We convert $\mathrm{CI}_{\mathrm{re}}$ to an LAI-scale proxy $\hat{L} = 0.5\,\mathrm{CI}_{\mathrm{re}}^{1.2}$ (falling back to an

NDVI-based Beer–Lambert estimate where the red-edge ratio is invalid), mask cloud-contaminated composites, and average over the county's crop pixels to obtain $\hat{L}_{c,y}(d)$ at each composite date $d \in \mathcal{D}_{c,y}$. Each candidate trajectory's simulated LAI, $L_{T_i}(d)$, is scored by the fraction of cloud-free composite dates on which it lies within a tolerance τ of the proxy, and the trajectory is retained if that score reaches an acceptance threshold ρ,

$$s(T_i) = \frac{1}{|\mathcal{D}_{c,y}|} \sum_{d \in \mathcal{D}_{c,y}} \mathbb{1}\big[\, |L_{T_i}(d) - \hat{L}_{c,y}(d)| \le \tau \,\big], \qquad \mathcal{T}^*_{c,y} = \{T_i \in \mathcal{T}_{c,y} \mid s(T_i) \ge \rho\}, \quad (1)$$

with $\tau = 2.0$ LAI units and $\rho = 0.6$; if no candidate reaches ρ, the three highest-scoring candidates are retained instead, so that every county-season contributes a teacher. The proxy is used solely for filtering candidate trajectories; supervision targets remain simulator-derived physiological variables rather than satellite observations. The value $\tau = 2.0$ is a practical uncertainty budget covering satellite retrieval error, county-level spatial aggregation, and structural uncertainty in the simulator, rather than a threshold tuned on downstream yield performance; Appendix B reports a sensitivity analysis over $\tau \in \{1.0, 1.5, 2.0, 2.5\}$. When multiple trajectories satisfy Equation 1, the corresponding process variables are averaged into a single consensus teacher trajectory for that county-season. On average, 10.2 of 12 candidate scenarios are retained for corn and 7.5 of 12 for soybean (Appendix B); the lower and more variable soybean retention is consistent with APSIM's tendency to overestimate peak LAI and delay senescence for indeterminate soybean varieties relative to the satellite proxy.

Figure 2 illustrates the resulting agreement between retained APSIM trajectories and satellite-observed canopy development for both corn and soybean. Because the same $\mathrm{CI}_{\mathrm{re}}$ proxy that drives selection is also used to assess agreement here, this comparison confirms that the filter behaves as intended rather than independently validating the physiological accuracy of the retained trajectories or of the non-LAI targets (biomass, water stress) that accompany them; we return to this limitation in Section 7.

4.4 Stage A: Process-Guided Continued Pretraining

The filtered teacher bank is used to continue training the pretrained SPEAR-NeXT temporal encoder q^m_θ for each modality m, together with its horizon-forecasting heads, on pixel sequences of the target crop drawn from the three teacher states. The SPEAR state encoders that produce the per-timestep spectral state $\boldsymbol{e}^m_t$ remain frozen throughout this stage (Section 3); only q^m_θ, its horizon heads, and a small auxiliary process head h_ϕ are trained.

For each observed timestep t, the causal contextual state $\boldsymbol{r}_t = q_\theta(\boldsymbol{e}_{1:t})$ is passed to h_ϕ, which predicts the P selected physiological variables at that timestep, $\hat{\boldsymbol{p}}_t = h_\phi(\boldsymbol{r}_t) \in \mathbb{R}^P$. Because $\boldsymbol{r}_t$ is causal, h_ϕ can use only observations up to t to predict the crop's state. Rather than supervising only an end-of-season value, the auxiliary loss is applied across every observed timestep up to the current training cutoff, so the model is pushed toward temporally consistent crop dynamics rather than relying on a single terminal target. Training minimizes

$$\mathcal{L}_A = \mathcal{L}_{\mathrm{NEP}} + \lambda\, \mathcal{L}_{\mathrm{proc}}, \qquad \mathcal{L}_{\mathrm{proc}} = \frac{1}{|\mathcal{M}|} \sum_{(n,t) \in \mathcal{M}} \sum_{p=1}^{P} \mathrm{Huber}\big(\hat{p}_{n,t,p},\, \tilde{p}_{c(n),y(n)}(d_{n,t})\big), \quad (2)$$

where $\mathcal{L}_{\mathrm{NEP}}$ is SPEAR-NeXT's own self-supervised forecasting objective (Appendix G), $\tilde{p}$ is the teacher value for pixel n's county-season at the composite date $d_{n,t}$ (matched to the nearest simulated day within ± 4 days), and $\mathcal{M}$ masks timesteps with no matched teacher. Teacher values are z-scored independently for each training cutoff, using only teacher observations up to that cutoff, so normalization statistics never depend on the future part of the season. λ is small (10^{-3}, or 5×10^{-4} for the soybean Sentinel-1 and climate encoders) so that the self-supervised forecasting objective continues to dominate training and the pretrained representation is refined rather than replaced (Appendix C reports the resulting representation similarity). Training uses AdamW (learning rate 10^{-4}, weight decay 0.05, 8 warm-up epochs, EMA decay 0.999, batch size 256, up to 50 epochs) with model selection on a held-out random 10% of pixel-seasons.

No yield label is used anywhere in stage A. A separate adapted encoder is trained for each crop and for each auxiliary target subset examined in Section 6.1; the subsets are nested (LAI alone; LAI and biomass; LAI, biomass, and water stress), so each is a distinct stage-A run rather than a shared

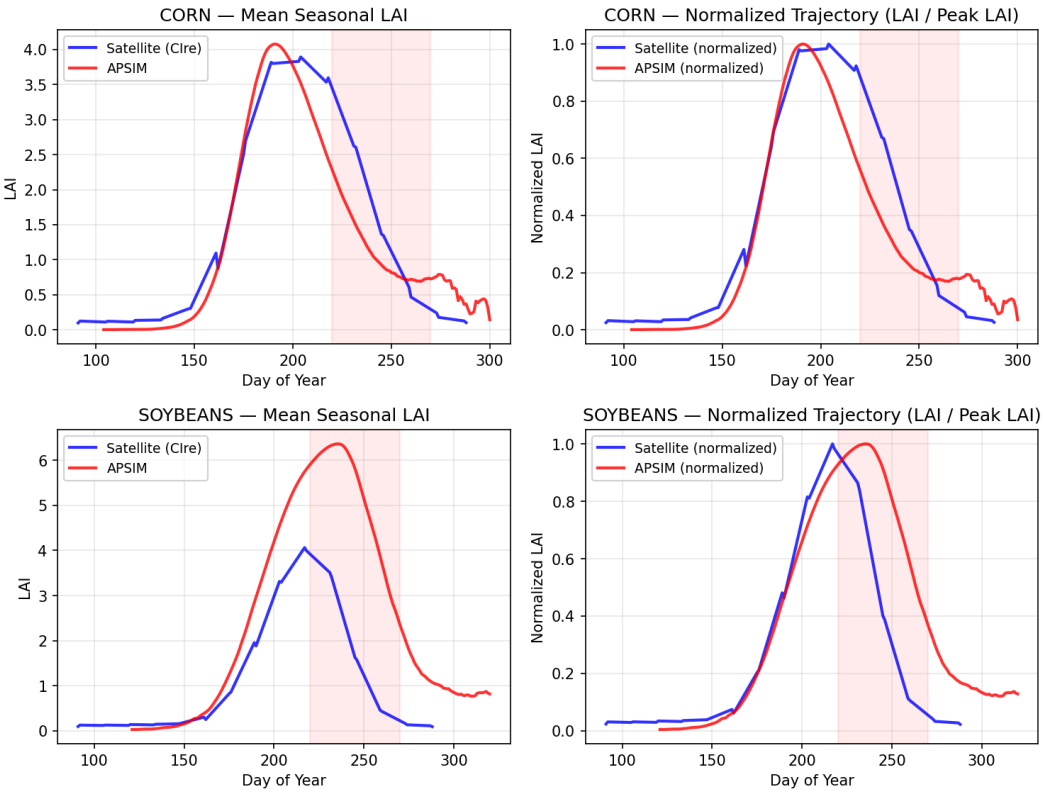


Figure 2: Agreement between retained APSIM trajectories and the Sentinel-2 canopy proxy for corn (top) and soybean (bottom), shown as mean seasonal LAI (left) and peak-normalized trajectories (right). The timing of the simulated and observed peaks aligns closely for corn; for soybean the satellite peak occurs earlier than the retained APSIM peak, consistent with APSIM's tendency to delay senescence for indeterminate varieties.

model with different heads enabled. After training, the process head h_ϕ is discarded; it plays no role in yield prediction.

Stage A uses satellite inputs and simulator outputs from all seasons of the three teacher states, analogous to SPEAR-NeXT's own self-supervised pretraining. Since no yield statistic is observed at this or any earlier stage, Section 4.5's evaluation tests generalization to unseen yields, though not unseen locations (Section 5).

4.5 Stage B: Yield Probing

The stage-A adapted encoders are frozen, and the 485-dimensional county feature $\boldsymbol{x}_{t_c}$ described in Section 3 is extracted at each forecasting cutoff t_c. A lightweight probe f_ψ (an MLP with hidden sizes 256 and 128, LayerNorm, GELU activations, and dropout 0.3) is trained with mean-squared error against standardized county yields of the training years, using AdamW (learning rate 10^{-3}, weight decay 10^{-2}, cosine learning-rate schedule, up to 200 epochs) with early stopping on a validation split drawn from the training years only. Neither simulator outputs nor the process head h_ϕ are used in this stage or at inference: prediction relies solely on satellite and climate observations.

The unadapted SPEAR-NeXT backbone is evaluated with the identical feature, probe, folds, and cutoffs, so stage-A adaptation is the only difference between the two conditions. We use a frozen-feature probe rather than end-to-end fine-tuning because it measures what the representation already encodes, and applies unchanged to every backbone compared, including those without a yield-specific head.

5 EXPERIMENTAL SETUP

5.1 BENCHMARK DATASET

Experiments are conducted on the AgroBench benchmark (Singh et al., 2026), a large-scale multimodal dataset for county-level crop monitoring and yield prediction (Section 3). The benchmark combines Sentinel-2 multispectral imagery, Sentinel-1 SAR observations, MODIS land-surface temperature, CHIRPS precipitation, SRTM elevation, and USDA county-level agricultural statistics across multiple growing seasons. The APSIM teacher-generation configuration is summarized in Section 4.2 and detailed in Appendix A; simulator-derived supervision is constructed only for Iowa, Illinois, and Indiana, where reliable region-specific agronomic management practices are available, while evaluation (below) extends to the full set of corn- and soybean-producing states in AgroBench.

5.2 PROTOCOL

We use leave-one-year-out (LOYO) evaluation over 2017–2024: for each held-out year the probe is trained on all other years and tested on all counties of the held-out year, and we report the mean over years of the per-year county-level R^2. Forecasts are issued at seven cutoffs (15 April, 13 May, 10 June, 8 July, 5 August, 2 September, 14 October), each using only composites up to the cutoff. Corn is evaluated on eight states and soybean on eleven (Section 3); teachers exist only for Iowa, Illinois, and Indiana, so the remaining states test whether adaptation benefits counties with no simulation.

Because stage-A adaptation uses no yield label, each crop's adapted encoder is trained once on all seasons of the three teacher states and reused across every held-out year in stage B; LOYO year-holdout therefore governs only the yield probe, not the representation itself. Furthermore, LOYO holds out years, not locations, so the probe sees other years of every evaluated state; we describe our results as evidence that simulator-guided *adaptation* transfers beyond the teacher region, not as evidence of spatial out-of-distribution generalization, which would require a state-disjoint split (Section 7).

5.3 BASELINES

(i) *Random forest* (Breiman, 2001) on hand-crafted county features (Sentinel-2 band means/std and NDVI/EVI summaries up to the cutoff), our only non-representation-learning baseline. (ii) *Presto* (Tseng et al., 2023) and (iii) *TESSERA* (Feng et al., 2026), pretrained EO models with embeddings computed up to the cutoff (partial-season TESSERA plus NDVI summaries, avoiding post-cutoff imagery). (iv) *SPEAR-NeXT* (Ranjan et al., 2026b): the same backbone from which stage A starts, with no stage-A adaptation applied.

All three foundation-model baselines are evaluated under the identical stage-B protocol used for SPEAR-SIM (Section 4.5): the same frozen-feature extraction, probe, and LOYO folds, with the probe the only yield-supervised component. The raw-feature random forest instead uses its own pipeline and is reported as a separate, non-representation reference.

5.4 IMPLEMENTATION

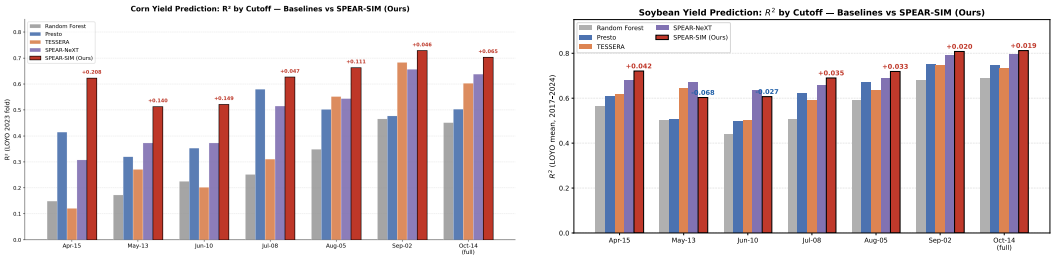
APSIM simulations use APSIM Next Generation (v2026.6.8052.0) via apsimNGpy; the foundation model and probes are implemented in PyTorch and trained on two NVIDIA RTX 3090 GPUs (Appendix F).

6 RESULTS

We evaluate the proposed framework using the LOYO protocol (Section 5). Although APSIM-derived teacher labels are generated only for Iowa, Illinois, and Indiana, the adapted representations are evaluated across eight states for corn and eleven for soybean without additional simulator supervision, assessing whether stage-A adaptation transfers beyond the teacher region. Teacher agreement with the satellite proxy is reported in Section 4.3 and Figure 2.

Table 1: Corn yield R^2 (LOYO mean, 2017–2024), eight states; teachers from Iowa, Illinois, and Indiana only. Best per cutoff in bold.

Model	Apr	May	Jun	Jul	Aug	Sep	Oct
Random forest (raw features)	0.149	0.173	0.225	0.252	0.349	0.466	0.451
Presto	0.415	0.320	0.353	0.580	0.502	0.477	0.503
TESSERA	0.121	0.271	0.202	0.311	0.552	0.683	0.603
SPEAR-NeXT (no simulator)	0.308	0.373	0.373	0.515	0.544	0.657	0.638
SPEAR-SIM (ours)	**0.623**	**0.513**	**0.522**	**0.627**	**0.663**	**0.729**	**0.703**



(a) Corn

(b) Soybean

Figure 3: Comparison of county-level yield prediction performance (R^2) across seasonal forecasting cutoffs for corn (a) and soybean (b). Annotations show the change from SPEAR-SIM relative to the strongest competing baseline at each cutoff, in red where SPEAR-SIM improves and in blue where it does not. SPEAR-SIM improves on the strongest baseline at every corn cutoff, but is below the unadapted SPEAR-NeXT backbone in May and June for soybean.

Simulator-guided adaptation versus the same backbone. Table 1 shows that SPEAR-SIM improves on the unadapted SPEAR-NeXT backbone at every corn forecasting cutoff, and also outperforms the raw-feature random forest, Presto, and TESSERA at every cutoff. The largest gain is in April (0.308 to 0.623), where only a few composites are available and the backbone must rely on how the season is starting; Figure 3 shows this gain narrowing but persisting through the season, reaching 0.729 in September. For soybean (Table 2), SPEAR-SIM is the best-performing model at every cutoff from July onward (0.690, 0.719, 0.808, and 0.812 for July through October, ahead of all four other models at each of those cutoffs), but it is *worse* than the unadapted SPEAR-NeXT backbone in May and June (0.603 vs. 0.671, and 0.607 vs. 0.634). Soybean canopy development and its APSIM representation are less reliable early in the season: teacher retention is lower than for corn (7.5 vs. 10.2 of 12 scenarios on average), and APSIM overestimates peak LAI and delays senescence relative to the satellite proxy (Section 4.3, Figure 2), so early-season soybean teachers are noisier.

Comparison with other foundation models. Presto is competitive early in the corn season (0.415 in April) and TESSERA late (0.683 in September), but neither matches SPEAR-SIM at any cutoff, for either crop. Because these baselines differ from SPEAR-NeXT in architecture, pretraining data, and feature construction, we treat these comparisons as context rather than controlled evidence; the SPEAR-NeXT row provides the controlled comparison (Section 4.5).

6.1 Which process variables help?

We train separate stage-A encoders with three nested target sets — LAI alone; LAI and biomass; and LAI, biomass, and water stress — and evaluate each with the identical stage-B protocol (Appendix D, Tables 9–10). Every target subset improves on the unadapted backbone at all seven corn cutoffs, so the corn improvement does not depend on a single choice of auxiliary targets, though LAI alone is strongest mid-season while the full combination is best at the earliest and latest cutoffs. For soybean, no single target subset dominates throughout the season. These results suggest that different physiological variables carry complementary information at different stages of crop

Table 2: Soybean yield R^2 (LOYO mean, 2017–2024), eleven states. SPEAR-SIM is below the unadapted backbone in May and June.

Model	Apr	May	Jun	Jul	Aug	Sep	Oct
Random forest (raw features)	0.563	0.503	0.438	0.507	0.592	0.681	0.690
Presto	0.609	0.507	0.497	0.620	0.671	0.748	0.746
TESSERA	0.618	0.645	0.503	0.589	0.633	0.745	0.734
SPEAR-NeXT (no simulator)	0.679	**0.671**	**0.634**	0.655	0.686	0.788	0.793
SPEAR-SIM (ours)	**0.721**	0.603	0.607	**0.690**	**0.719**	**0.808**	**0.812**

development, and that the relative value of each depends on both crop and growth stage, rather than following a single additive trend.

6.2 What changes in the representation?

Comparing county embeddings before and after stage A shows that stage A refines rather than replaces the pretrained representation (Appendix C): representation similarity is high in aggregate (linear CKA above 0.92 for both crops) despite substantial per-county refinement, and t-SNE projections show a smoother seasonal progression after adaptation. This describes how the representation changes; it does not by itself establish physiological accuracy, which we return to in Section 7.

7 Discussion and limitations

What the evidence supports. The controlled comparison shows that simulator-derived process supervision during continued self-supervised pretraining yields representations more predictive of yield, especially early in the corn season and across states without simulator supervision. We do not claim the adapted representation is physiologically *correct* — no independent field measurements of LAI, biomass, or water stress are available at this scale — and instead call it *simulator-consistent*, using downstream yield prediction as the evidence for its usefulness.

Teacher validity. Teacher selection uses the same canopy proxy against which Figure 2 compares the retained teachers, so that agreement is partly by construction: it confirms the selection rule behaves as intended, not that the trajectories are independently accurate. Biomass and water-stress teachers inherit validity only through the LAI filter and APSIM's internal consistency, and the chlorophyll-index-to-LAI mapping is empirical and uncalibrated for these counties; soybean teachers are the least reliable, consistent with the weaker early-season soybean results (Section 6).

Scope. Teachers exist for three states and two crops, with APSIM configurations chosen per region rather than calibrated against field data (Appendix A). Each stage-A encoder is trained once with a single seed, so year-to-year stability beyond the per-year means in Section 6 is not established. Evaluation holds out years, not locations, so results support transfer of the *adaptation* beyond the teacher region, not spatial out-of-distribution generalization; a state-disjoint split, training both stages only on the teacher states and testing on the rest, would be needed for that stronger claim. Natural extensions include such a split, alternative simulators such as DSSAT (Jones et al., 2003), retaining the APSIM ensemble as a distribution rather than a mean, and adaptively weighting targets by crop.

8 Conclusion

SPEAR-SIM uses mechanistic crop simulations, filtered by satellite observations, as training-time-only supervision for continued pretraining of a temporal EO foundation model. With no yield labels during adaptation and no simulator at inference, it improves county yield forecasting over the same backbone at every corn cutoff—most strongly in April—and at five of seven soybean cutoffs, including states beyond the three used for teacher generation. The approach generalizes beyond APSIM and agriculture: observation-constrained teacher selection, paired with label-free adaptation followed by label-based probing, offers a general recipe for bringing mechanistic simulators into foundation-model adaptation wherever observations can validate a simulator's output.

ETHICS STATEMENT

All data are public: Sentinel-1/2, MODIS, CHIRPS, SRTM, NASA POWER, SSURGO and USDA NASS county statistics, which are aggregated and contain no personal information. No human subjects are involved. Yield forecasts can influence markets and policy; our models are research artefacts evaluated on historical data and should not be used operationally without independent validation. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Section 4 and Appendices A–F give the teacher-generation configuration, selection rule, stage-A objective and all hyperparameters. The anonymized supplementary code contains weather and soil acquisition, the APSIM batch simulation, teacher selection, the stage-A training scripts with the process head, the stage-B evaluation harness and the raw-feature baseline. The SPEAR-NeXT backbone is described in Ranjan et al. (2026b); its implementation is not part of this submission, and the supplement includes a documented interface stub specifying every call made to it. AgroBench is described in Singh et al. (2026). Adapted checkpoints and teacher files will be released upon publication.

AI USE STATEMENT

A large language model was used to assist with restructuring and editing the manuscript, preparing the LaTeX source and the overview figure, and reviewing and writing parts of the evaluation code. All technical content, experiments and numbers were produced, checked and are taken responsibility for by the authors.

REFERENCES

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, et al. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint*, 2025.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

J. G. P. W. Clevers and A. A. Gitelson. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation*, 23:344–351, 2013.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211, 2022.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34818–34831, 2026.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, pp. 5506–5538, 2023.

Erhu He, Yiqun Xie, Licheng Liu, Weiye Chen, Zhenong Jin, and Xiaowei Jia. Physics guided neural networks for time-aware fairness: An application in crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14223–14231, 2023.

Dean Holzworth, Neil I. Huth, Justin Fainges, Hamish Brown, Eric Zurcher, Rogerio Cichota, Steve Verrall, Neville I. Herrmann, Bangyou Zheng, and Val Snow. APSIM next generation: Overcoming challenges in modernising a farming systems model. *Environmental Modelling & Software*, 103:43–51, 2018.

Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan Read, Jacob Zwart, Michael Steinbach, and Vipin Kumar. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 558–566, 2019.

J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A. Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie. The DSSAT cropping system model. *European Journal of Agronomy*, 18(3):235–265, 2003.

Michiel Kallenberg, Dilli Paudel, Ioannis N. Athanasiadis, et al. CY-Bench: A comprehensive benchmark dataset for sub-national crop yield forecasting. *Earth System Science Data*, 18(6): 3997–4018, 2026.

Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, Drew Gholson, Nicolas Ashwell, Tri Setiyono, Brenda Tubana, Lu Peng, Magdy Bayoumi, and Nian-Feng Tzeng. MMST-ViT: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5774–5784, 2023.

Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nian-Feng Tzeng. An open and large-scale dataset for multi-modal climate change-aware crop yield predictions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5375–5386, 2024.

Licheng Liu, Shaoming Xu, Jinyun Tang, Kaiyu Guan, Timothy J. Griffis, Matthew D. Erickson, Alexander L. Frie, Xiaowei Jia, Taegon Kim, Lee T. Miller, Bin Peng, Shaowei Wu, Yufeng Yang, Wang Zhou, Vipin Kumar, and Zhenong Jin. KGML-ag: A modeling framework of knowledge-guided machine learning to simulate agroecosystems: A case study of estimating N_2O emission using data from mesocosm experiments. *Geoscientific Model Development*, 15(7):2839–2858, 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Rajiv Ranjan, Udaiveer Singh, Anjali Aggarwal, Shashank Tamaskar, and Dharmendra Saraswat. SPEAR: Self-supervised sample efficient pixel-level multi-modal spectral fusion for earth observation applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 1445–1454, 2026a.

Rajiv Ranjan, Udaiveer Singh, Shashank Tamaskar, and Dharmendra Saraswat. SPEAR-NeXT: Causal latent forecasting across multiple horizons for spectral–temporal earth representation learning. *arXiv preprint arXiv:2609.16871*, 2026b.

Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V. Archontoulis. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports*, 11(1):1606, 2021.

Udaiveer Singh, Rajiv Ranjan, Shashank Tamaskar, and Dharmendra Saraswat. AgroBench: A reproducible multimodal benchmark for weakly supervised crop yield learning from county statistics and pixel observations. *arXiv preprint arXiv:2609.26809*, 2026.

Adam H. Sparks. nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R. *Journal of Open Source Software*, 3(30):1035, 2018.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

USDA Natural Resources Conservation Service. Soil survey geographic (SSURGO) database, 2024.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Caleb A. Winn, Sotirios Archontoulis, and Jode Edwards. Calibration of a crop growth model in APSIM for 15 publicly available corn hybrids in North America. *Crop Science*, 63(2):511–534, 2023.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, and Stefano Ermon. SustainBench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep Gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Runlong Yu, Chonghao Qiu, Robert Ladwig, Paul Hanson, Yiqun Xie, and Xiaowei Jia. Physics-guided foundation model for scientific discovery: An application to aquatic science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28548–28556, 2025.

A APSIM CONFIGURATION

Simulations use APSIM Next Generation v2026.6.8052.0 driven through apsimNGpy, with parameters written directly into the simulation JSON. Weather and soil are fixed within a county-season, so ensemble members differ only in management. Counties are assigned to latitude zones per state (Table 3); cultivar and candidate sowing dates follow the zone (Tables 4–5), and nitrogen is applied at three levels (corn 120/160/200, soybean 0/10/20 kg N ha^{-1}).

Rather than using a single cultivar across the study region, cultivar selection follows latitude-dependent maturity recommendations to better reflect regional agronomic practice: progressively longer-maturity cultivars are used in southern zones and shorter-maturity cultivars at higher latitudes (Figure 4). This reduces systematic bias from an unrealistic single cultivar choice while letting the simulator capture regional differences in crop development. Daily weather forcing comes from NASA POWER (Sparks, 2018), including maximum and minimum temperature, precipitation, and incoming solar radiation; weather and soil inputs are held fixed across all management scenarios within a county-season, so differences between simulated trajectories arise only from management uncertainty, not environmental variability.

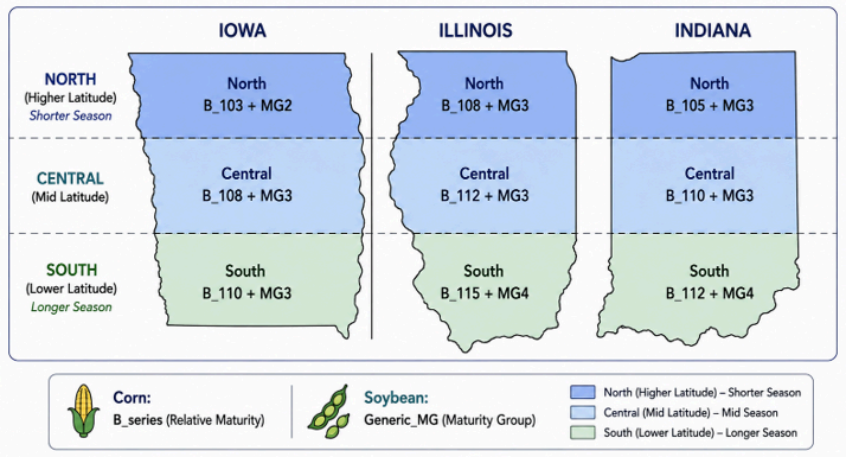


Figure 4: Latitude-matched agronomic zones used for cultivar assignment in APSIM simulations. Corn hybrids and soybean maturity groups are selected according to geographic latitude to reflect realistic regional growing conditions.

For corn, template soils gave better canopy agreement than SSURGO (agreement score 0.920 vs. 0.849; 0.856 after correcting drainage-rate and layer transfer), because lower drained-upper-limit values in SSURGO profiles produced early water stress and premature senescence relative to the county-aggregated satellite observations (Table 6). For soybean we retained SSURGO profiles despite a similar structural difference, because reproductive-stage water stress is itself a meaningful supervision target for that crop.

Table 3: Latitude zones.

State	North	Central	South
Iowa	$\geq 42.5°$N	41.0–42.5°N	$< 41.0°$N
Illinois	$\geq 41.0°$N	39.0–41.0°N	$< 39.0°$N
Indiana	$\geq 40.5°$N	39.0–40.5°N	$< 39.0°$N

For both crops, cultivar selection is fixed within each latitude zone and is not varied during ensemble generation; management uncertainty is introduced only through the combination of sowing date and nitrogen rate, giving twelve scenarios per county-season (twenty-four for the northernmost Iowa zone, which uses two calibrated soybean cultivars). Across both crops, the resulting corpus contains 23,592 mechanistic crop simulations spanning 934 corn and 929 soybean county-seasons. Although simulations are generated only in these three states, the adapted encoders are subsequently evaluated

Table 4: Corn management scenarios (cultivar and sowing dates per zone).

Zone	Cultivar	Sowing dates
IA N	B_103	Apr 25, May 5, May 15, May 25
IA C	B_108	Apr 20, May 1, May 10, May 20
IA S	B_110	Apr 18, Apr 28, May 8, May 18
IL N	B_108	Apr 25, May 5, May 15, May 25
IL C	B_112	Apr 20, Apr 30, May 10, May 20
IL S	B_115	Apr 5, Apr 15, Apr 25, May 5
IN N	B_105	Apr 27, May 5, May 13, May 20
IN C	B_110	Apr 20, Apr 28, May 6, May 15
IN S	B_112	Apr 13, Apr 21, Apr 29, May 8

Table 5: Soybean management scenarios.

Zone	Cultivar	Sowing dates
IA N	MG2	May 5, May 15, May 25, Jun 5
IA C, IA S, IL C	MG3	May 1, May 10, May 20, Jun 1
IL N	MG3	May 5, May 15, May 25, Jun 5
IL S, IN S	MG4	Apr 25, May 5, May 15, May 25
IN N	MG3	May 10, May 20, Jun 1, Jun 10
IN C	MG3	May 5, May 15, May 25, Jun 5

across the larger multi-state AgroBench footprint (Section 5) without generating additional simulator outputs there.

B Observation-constrained teacher selection: statistics and sensitivity

Direct field observations of LAI, biomass, or water stress are unavailable at the scale needed here, so canopy development is instead estimated from the Sentinel-2 red-edge chlorophyll index $\mathrm{CI}_{\mathrm{re}}$ (Section 4.3), aggregated into the same fourteen-day composites used elsewhere in AgroBench. This proxy is used only to select physically consistent APSIM trajectories; it is never itself a supervision target. Unlike data-assimilation approaches that adjust a simulator's internal state to match observations, our procedure only identifies which unmodified candidate simulations are consistent with what was observed and discards the rest.

Observation-constrained filtering is applied independently to every county-season after the full management ensemble has been generated; the number of retained scenarios is stored alongside each teacher record to keep the procedure fully auditable.

C Representation analysis

To quantify how stage-A adaptation modifies the pretrained representation space, we compare frozen and adapted county embeddings using cosine similarity and linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Cosine similarity measures the magnitude of change for individual county-season embeddings; CKA measures whether the global geometry of the representation space is preserved.

Mean cosine similarities well below 1 indicate that individual county embeddings undergo substantial refinement during adaptation, while CKA values above 0.92 indicate that the overall geometry of the pretrained representation space is largely preserved: stage A appears to refine, rather than replace, the pretrained representation.

Figure 8 visualizes county embeddings before and after stage-A adaptation using t-SNE projections colored by composite date. For corn, adapted embeddings show a clearer progression through the growing season while preserving temporal continuity between adjacent composites; soybean shows

Table 6: Diagnostic experiments used to determine the soil parameterization for corn APSIM simulations.

Run	Configuration	Corn LAI score
D1	Template soil	0.920
D3	SSURGO	0.849
D7	SSURGO + SWCON fix + layer fix	0.856

Table 7: Observation-constrained selection statistics ($\tau = 2.0$, $\rho = 0.6$).

	Corn	Soybean
County-seasons with teachers	934	929
Scenarios per county-season	12	12 (24 for IA N)
Total APSIM simulations	11,208	12,384
Retained scenarios, mean (sd)	10.22 (2.11)	7.54 (5.78)
Retained, median [min, max]	12 [2, 12]	6 [1, 24]
County-seasons retaining all scenarios	474 (50.7%)	134 (14.4%)
County-seasons retaining ≥ 9	840 (89.9%)	375 (40.4%)

a similar, though noisier, organization, consistent with its greater indeterminate-growth variability. This is a qualitative observation about the organization of the representation, not independent evidence of physiological accuracy (Section 7).

D STAGE-A TARGET ABLATION: DETAILED RESULTS

Tables 9–10 report per-cutoff R^2 for stage-A encoders trained with each nested target subset (Section 6.1), each row a separately trained stage-A encoder evaluated with the identical stage-B probe. For corn, the relative contribution of each auxiliary variable varies across forecasting cutoffs: LAI supervision alone provides the strongest signal in the middle of the season (June through August), while the full combination of LAI, biomass, and water stress is best at the earliest and latest cutoffs. For soybean, no single target subset dominates throughout the season: LAI alone is strongest in May, August, and October, LAI plus biomass is strongest in June and July, and the full combination is strongest in April and September.

E SENSITIVITY TO THE FILTERING THRESHOLD

The canopy agreement threshold τ controls the trade-off between teacher quality and ensemble diversity: smaller values retain fewer simulations while enforcing stronger agreement with satellite observations, whereas larger values preserve more management variability at the cost of admitting less consistent trajectories. Table 11 summarizes this trade-off across four candidate values.

As expected, increasing τ monotonically increases both the agreement score and the number of retained management scenarios, since a larger tolerance admits a larger fraction of the simulated ensemble. Although $\tau = 2.5$ reaches slightly higher agreement scores than $\tau = 2.0$, it also admits simulations whose deviation from the proxy exceeds the uncertainty budget the threshold is meant to represent. We therefore treat $\tau = 2.0$ LAI units as a principled operating point rather than a value empirically optimized for downstream performance: it represents the combined uncertainty arising from satellite retrieval error, county-scale spatial aggregation, APSIM's own structural uncertainty, and the cultivar and soil parameterizations described in Appendix A, while retaining enough management diversity for the consensus teacher to remain meaningful. All experiments reported in the main text use $\tau = 2.0$.

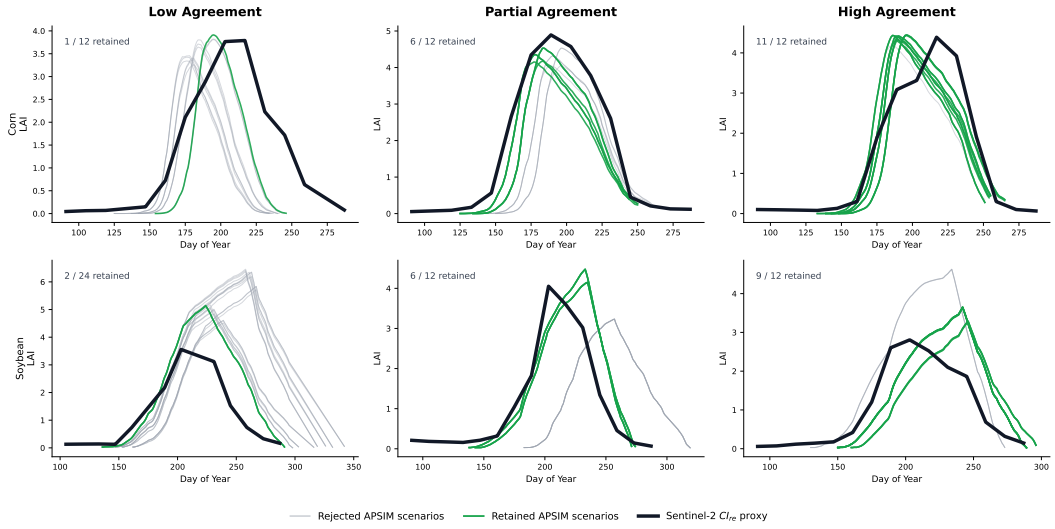


Figure 5: Examples of the observation-constrained selection procedure at low, partial, and high agreement, for corn (top) and soybean (bottom). Gray: rejected candidate trajectories; green: retained; black: Sentinel-2 $\mathrm{CI}_{\mathrm{re}}$ canopy proxy. Rejected trajectories typically show unrealistic emergence timing, peak canopy development, or senescence relative to the proxy; retained trajectories preserve management-driven variability while tracking the observed season.

Table 8: Representation similarity before and after stage-A adaptation.

Crop	Mean cosine	Median cosine	Linear CKA
Corn	0.124	0.260	0.9247
Soybean	0.183	0.265	0.9245

F HYPERPARAMETERS

G SPEAR-NEXT PRETRAINING OBJECTIVE

For reference, SPEAR-NeXT's self-supervised forecasting objective (not a contribution of this paper; described in full in Ranjan et al. (2026b)) is

$$\mathcal{L}_{\mathrm{NEP}} = \sum_{k=1}^{K} w_k \Big[\alpha \left(1 - \cos(\hat{\boldsymbol{e}}_{t+k|t}, \boldsymbol{e}_{t+k})\right) + (1-\alpha)\, \tfrac{1}{D} \|\hat{\boldsymbol{e}}_{t+k|t} - \boldsymbol{e}_{t+k}\|_2^2\Big], \qquad w_k \propto k^{-1}, \quad (3)$$

where $\alpha \in [0, 1]$ balances a directional (cosine) term against a magnitude (squared-error) term, w_k down-weights longer horizons, and the sum is averaged over valid context–target pairs with stop-gradient targets. This is the objective used to pretrain the SPEAR-NeXT backbone (Section 3) and, unchanged, the first term of the stage-A objective $\mathcal{L}_A$ (Equation 2).

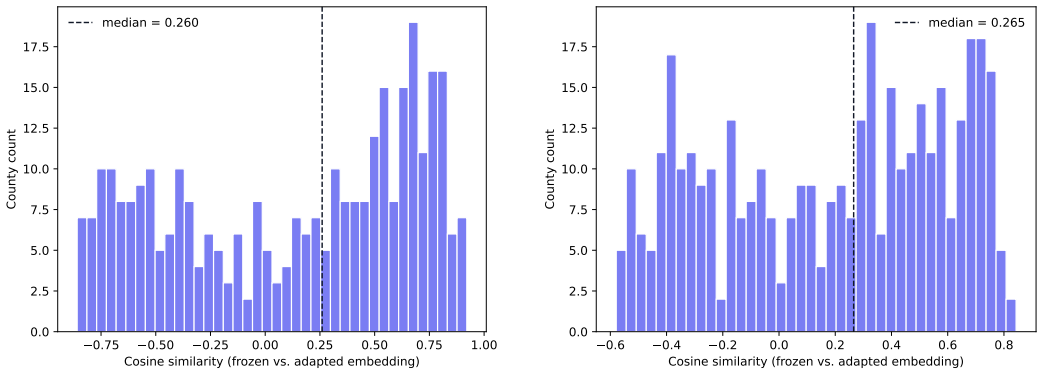


Figure 6: Distribution of cosine similarity between frozen and stage-A-adapted county embeddings for corn (left) and soybean (right).

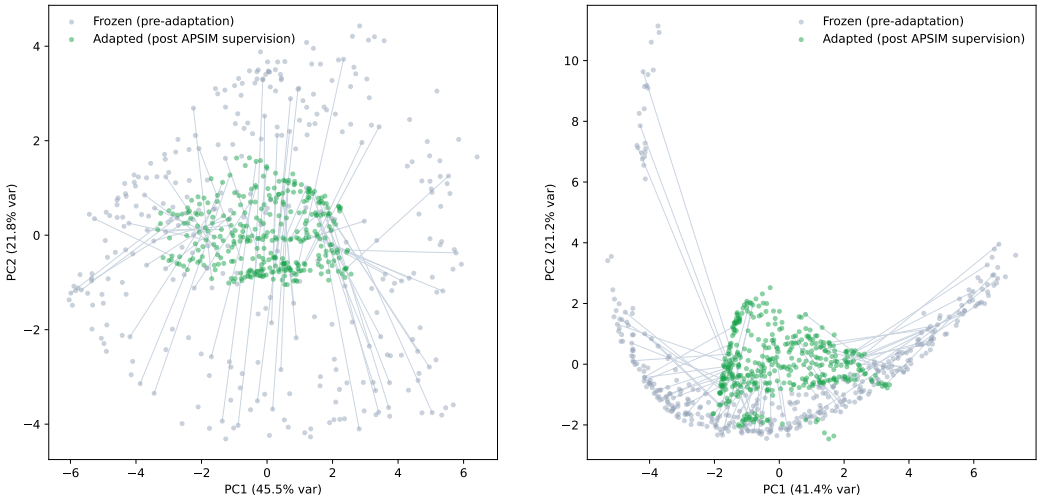


Figure 7: Principal-component projections of county embeddings before and after stage-A adaptation, for corn (left) and soybean (right). Adaptation refines local representations while preserving the overall geometry of the pretrained embedding space.

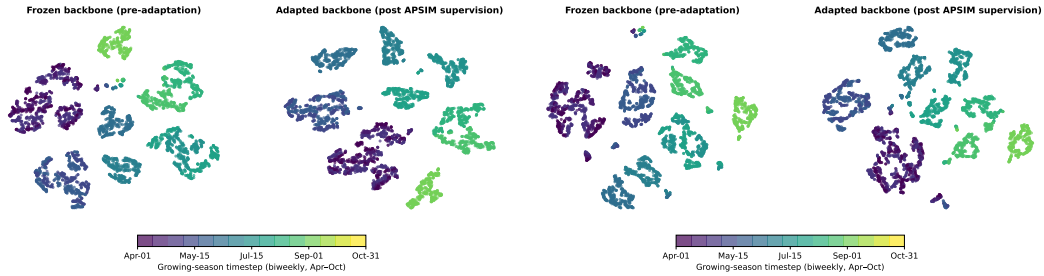


Figure 8: t-SNE projections of county embeddings before and after stage-A adaptation, colored by composite date, for corn (left) and soybean (right).

Table 9: Stage-A target ablation for corn (R^2, LOYO). Best per cutoff in bold.

Stage-A targets	Apr	May	Jun	Jul	Aug	Sep	Oct
none (SPEAR-NeXT)	0.308	0.373	0.373	0.515	0.544	0.657	0.638
LAI	0.580	0.489	**0.522**	**0.656**	**0.702**	0.703	0.687
LAI + biomass	0.604	0.476	0.497	0.632	0.668	0.724	0.702
LAI + biomass + water stress	**0.623**	**0.513**	**0.522**	0.627	0.663	**0.729**	**0.703**

Table 10: Stage-A target ablation for soybean (R^2, LOYO). Best per cutoff in bold.

Stage-A targets	Apr	May	Jun	Jul	Aug	Sep	Oct
none (SPEAR-NeXT)	0.679	**0.671**	**0.634**	0.655	0.686	0.788	0.793
LAI	0.713	0.665	0.592	0.688	**0.730**	0.806	**0.823**
LAI + biomass	0.692	0.603	0.620	**0.695**	0.711	0.776	0.812
LAI + biomass + water stress	**0.721**	0.603	0.607	0.690	0.719	**0.808**	0.812

Table 11: Sensitivity of observation-constrained teacher generation to the canopy agreement threshold (τ). Agreement is the canopy-agreement score of the consensus teacher with the satellite proxy. Values for $\tau \in \{1.0, 1.5, 2.5\}$ are from a calibration study; $\tau = 2.0$ is the production corpus used throughout the paper.

τ	Corn agreement	Soybean agreement	Retained (corn)	Retained (soy)
1.0	0.816	0.508	7.4	4.0
1.5	0.920	0.604	9.4	5.5
2.0	0.963	0.691	10.22	7.54
2.5	0.985	0.760	10.6	10.3

Table 12: Stage A (process-guided continued pretraining) and stage B (yield probe).

	Stage A	Stage B
Trainable	temporal encoder, horizon heads, process head	probe MLP
Frozen	SPEAR state encoders	SPEAR encoders, adapted temporal encoder
Data	target crop, Iowa/Illinois/Indiana, all seasons	target crop, 8/11 states, LOYO training years
Labels	APSIM teachers (no yield)	county yield (no simulator)
Loss	$\mathcal{L}_{\text{NEP}} + \lambda$ Huber	MSE on standardized yield
λ	10^{-3} (5×10^{-4} soybean S1/climate)	–
α in Eq. 3	0.5 (S2, climate), 0.2 (S1)	–
Optimizer	AdamW, lr 10^{-4}, wd 0.05, β=$(0.9, 0.95)$	AdamW, lr 10^{-3}, wd 10^{-2}
Schedule	8 warm-up epochs, $\leq$50 epochs, EMA 0.999	cosine, $\leq$200 epochs
Batch	256 pixel-seasons	128 county-seasons
Selection	10% random pixel-seasons, patience 15	validation split of training years, patience 20
Teacher alignment	nearest simulated day within ±4 days	–
Architecture	4 layers, width 128, 4 heads, 46 horizon heads	hidden (256, 128), dropout 0.3
Seed	42	42