PHYSCROP-RISK: PROGRESS-ALIGNED SEQUENTIAL RISK CONTROL FOR EARLY CROP MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Early crop monitoring requires repeated decisions from irregular observations before harvest. Two challenges arise: calendar time poorly reflects crop development, and repeated decisions accumulate false alarms even when each is calibrated separately. We present PhysCrop-Risk, a progress aligned framework that uses growing degree days (GDD) to define causal observation boundaries and decision checkpoints, aligns Sentinel-2 and weather observations on a common thermal scale, and calibrates low yield evidence over the full season. At three crop specific checkpoints, risk scores are converted into empirical one sided p-values against held out normal yield fields, and a shared significance budget controls the probability of at least one false watch. Two consecutive crossings define a more conservative action tier. On 2,171 fields spanning four countries, four crops, and nine seasons, the final checkpoint on a 322 field audit reaches an AUROC of 0.8331 and an AUPRC of 0.6245 at a median lead time of 62.5 days before harvest. Under the frozen protocol, watch and persistent-action tiers achieve field season false positive rates of 5.24% and 2.02%, with the watch confidence interval covering the 5% budget. An exploratory post hoc calibration analysis yields 4.84% and 1.61% but does not inherit the finite sample guarantee. These results show that progress aware alignment and sequence level calibration can turn partial season observations into reliable early crop risk decisions.

1 INTRODUCTION

Many real world monitoring problems require decisions from observations that arrive gradually before the final outcome is known. Early crop monitoring is one example: satellite imagery and weather accumulate throughout the season, while fields at risk of low yield may need to be identified well before harvest. Recent work has improved field scale and within season yield prediction from satellite, weather, and environmental observations (Xiao et al., 2024; Xiong et al., 2024; Zare et al., 2024; Mena et al., 2025). Early warning adds a sequential requirement: each decision must use only the observations available at that point, and reliability must hold as evidence accumulates.

Calendar time is an imperfect coordinate for crop development because growth rates vary across crops, locations, weather, and seasons. Growing degree days provide an established measure of thermal progress (McMaster & Wilhelm, 1997; Bonhomme, 2000), while satellite phenology has been used to recover crop timing and organize observations across growing conditions (Zhou et al., 2024; Duan et al., 2024). Progress therefore provides a more meaningful basis for comparing partial observations and defining early decision points.

Repeated decisions create a second difficulty: every checkpoint introduces another opportunity for a false alarm. This connects early monitoring to familywise error control and error spending over repeated tests (Tian & Ramdas, 2021; Lan & DeMets, 1983). The relevant error is therefore whether a normal field is flagged at least once over the season, motivating a shared false alert budget across sequential decisions.

We present PhysCrop-Risk, a progress-aligned framework for early crop risk monitoring from partial multimodal observations. A unified GDD coordinate defines observation boundaries, temporal alignment, and decision checkpoints. At three crop-specific checkpoints, Sentinel-2, weather, and environmental data jointly estimate final yield. Domain-normalized risk scores are calibrated against held-out normal-yield fields via empirical one-sided p-values under a shared significance

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.

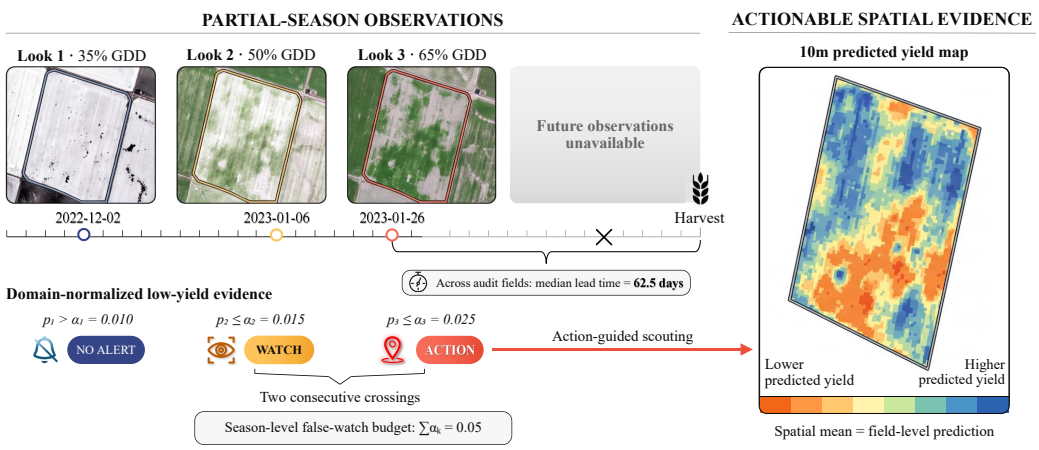

Figure 1: **PhysCrop-Risk: progress aligned, calibrated crop risk decisions from partial season observations.** GDD checkpoints set when evidence is evaluated. One crossing raises a watch, two consecutive crossings trigger an action under a season level false watch budget. A mean preserving 10 m map adds scouting evidence without changing the field prediction. Final checkpoint: median 62.5 days before harvest.

budget. Two consecutive threshold crossings trigger conservative alerts; a mean-preserving 10 m residual map captures within-field variability without altering field-level predictions (Figure 1). On 2,171 fields across four countries, four crops, and nine seasons, PhysCrop-Risk achieves accurate early yield forecasting and low-yield discrimination, with the final checkpoint a median of 62.5 days pre-harvest. Under the frozen protocol, watch and persistent-action tiers match the predefined field-season false positive budget. Post-hoc calibration enables a more aggressive regime but loses finite-sample guarantees.

Our contributions are fourfold:

- We formulate early crop monitoring as progress indexed sequential risk recognition, coupling causal partial observations with a field season error criterion for repeated decisions.
- We introduce a progress aligned multimodal framework in which GDD jointly defines temporal alignment and decision checkpoints across crops, regions, and seasons.
- We develop sequence level calibration that converts domain normalized predictions into empirical null evidence and controls the probability of a false watch over the monitored season under exchangeability.
- We connect calibrated field decisions to a mean preserving 10 m spatial refinement and evaluate the framework across held out fields, farms, seasons, temporal alignments, and within field yield maps.

2 Related Work

2.1 Learning from Partial and Asynchronous Time Series

Irregular sampling, missing observations, and heterogeneous inputs are central challenges in temporal learning. Continuous time and attention based models address irregular observations (Kidger et al., 2020; Shukla & Marlin, 2021), while Earth observation methods further handle missing acquisitions, heterogeneous modalities, and varying resolutions through multimodal fusion and large scale pretraining (Astruc et al., 2024; 2025; Do et al., 2025; Nedungadi et al., 2024; Tseng et al., 2025). Masked and temporal representation learning provide complementary approaches, including MAESTRO (Labatie et al., 2026) and TESSERA (Feng et al., 2026). In agriculture, satellite, weather, and environmental observations support multimodal yield prediction (Xiong et al., 2024; Mena et al., 2025; Miranda et al., 2026), including early and within season forecasting from partial

observations (Ven et al., 2026; Zare et al., 2024). PhysCrop-Risk extends this setting from partial sequence prediction to repeated decisions over a growing causal prefix.

2.2 Progress Aware Temporal Learning

Calendar time imperfectly reflects crop state because development varies with temperature, crop, location, and season. Thermal time and growing degree days provide established measures of crop development (McMaster & Wilhelm, 1997; Bonhomme, 2000). Recent Earth observation models exploit seasonal structure through vegetation forecasting, temporal spectral embeddings, and multitemporal reconstruction (Benson et al., 2024; Feng et al., 2026; Labatie et al., 2026), while agricultural models incorporate phenology into temporal representation and yield prediction (Li et al., 2026; Zhou et al., 2024; Luo et al., 2026). Thermal time has also been used to organize Sentinel-2 and weather trajectories for yield loss analysis (Duan et al., 2024). PhysCrop-Risk uses the same GDD checkpoints to define the causal prefix, temporal alignment, and decision time, so fields are evaluated at comparable stages of thermal development.

2.3 Calibration for Sequential Decisions

Conformal prediction provides finite sample calibration under exchangeability (Shafer & Vovk, 2008), while conformal risk control extends this principle to application level losses (Angelopoulos et al., 2024a). Recent work addresses nonexchangeable and multi source settings (Farinhas et al., 2024; Liu et al., 2024), subpopulation calibration (Shi et al., 2024; van der Laan & Alaa, 2024), and downstream decision support (Cresswell et al., 2024). Online conformal methods adapt calibration as observations arrive (Angelopoulos et al., 2024b; Gibbs & Candès, 2021; Podkopaev et al., 2024), with related work on structured time series outputs (Chen et al., 2024; Xu et al., 2024). Repeated testing further connects to familywise error control and error spending (Tian & Ramdas, 2021; Lan & DeMets, 1983). PhysCrop-Risk applies this sequence level view to crop monitoring, using checkpoint specific empirical nulls and a shared significance budget to control the probability of at least one false watch over the field season under exchangeability.

3 Problem Formulation

We consider early risk monitoring from observations that accumulate over a growing season. For field i, let y_i denote final yield, d_i its country and crop domain, and

$$X_i = \{x_{i,t}\}_{t=1}^{T_i}$$

its irregular multimodal observation sequence, together with static context u_i. The outcome y_i is observed only after harvest, while decisions must be made earlier from partial information.

We associate each observation with a monotone process coordinate $g_i(t)$. For crop c, let

$$L_{c,1} < \cdots < L_{c,K}$$

denote K progress checkpoints. The information available to field i at checkpoint k is the causal prefix

$$X_i^{(k)} = \{x_{i,t} \in X_i : g_i(t) \le L_{c_i,k}\}. \tag{1}$$

Thus $X_i^{(1)} \subseteq \cdots \subseteq X_i^{(K)}$, and each decision uses only information available at the corresponding stage of crop development.

Low yield is defined relative to a domain specific threshold τ_d,

$$z_i = \mathbb{1}[y_i \le \tau_{d_i}]. \tag{2}$$

At each checkpoint, the current causal prefix produces a low yield score $s_{i,k}$ and a watch decision $W_{i,k} \in \{0, 1\}$. Since the same field is evaluated repeatedly, the relevant error is whether a normal yield field is ever flagged during the season. We therefore seek decision rules satisfying

$$\Pr(\exists k \le K : W_{i,k} = 1 \mid z_i = 0) \le \alpha, \tag{3}$$

where α is a prespecified field season false watch budget.

PhysCrop-Risk instantiates $g_i(t)$ with accumulated growing degree days and combines progress aligned prediction with calibration across the full sequence of checkpoints.

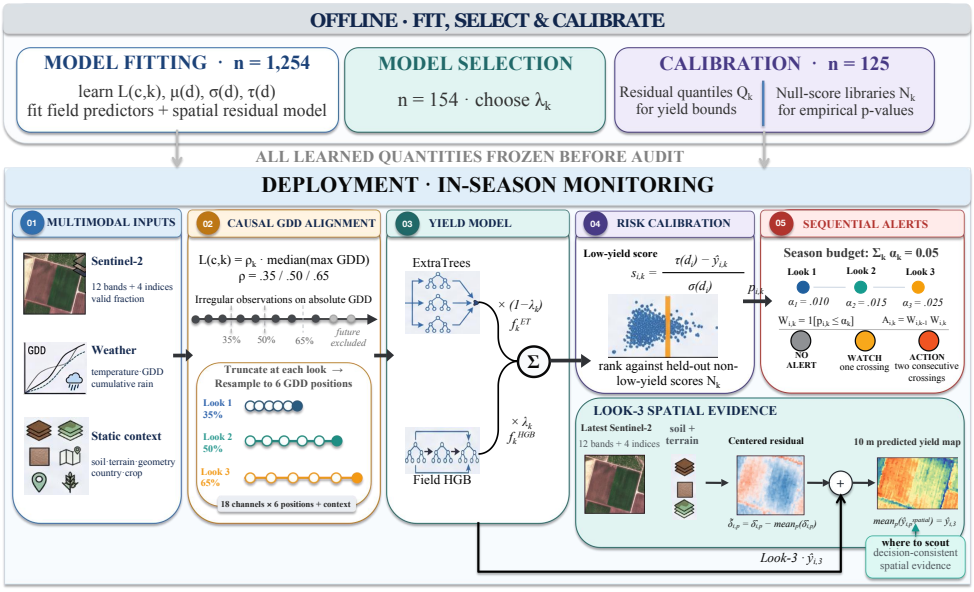

Figure 2: **PhysCrop-Risk framework.** Offline development freezes progress checkpoints, predictors, and calibration references before audit. In season, GDD aligns partial multimodal observations into field level yield estimates and normalized low yield evidence, calibrated across three checkpoints under a shared season false watch budget. A mean preserving spatial branch adds 10 m scouting evidence without changing the field level prediction.

4 METHOD

Figure 2 summarizes the framework. PhysCrop-Risk aligns partial observations by crop progress, converts yield estimates into domain normalized risk evidence, and calibrates repeated decisions under a shared field season error budget. A spatial refinement branch provides local evidence while preserving the field prediction.

4.1 PROGRESS ALIGNED MULTIMODAL PREDICTION

We use accumulated growing degree days (GDD) as the progress coordinate. For field i,

$$\Delta g_{i,u} = \mathrm{clip}\left(T_{i,u}^{\mathrm{mean}} - 5, 0, 25\right), \qquad g_i(t) = \sum_{u \le t} \Delta g_{i,u}. \tag{4}$$

Three crop specific checkpoints are estimated using only the model fitting partition and are then frozen:

$$L_{c,k} = \rho_k \, \mathrm{median}_{i \in D_{\mathrm{fit}},\, c_i = c} \left[\max_t g_i(t)\right], \qquad (\rho_1, \rho_2, \rho_3) = (0.35, 0.50, 0.65). \tag{5}$$

At checkpoint k, observations beyond $L_{c_i,k}$ are discarded and the remaining dynamic variables are interpolated on a fixed grid. The resulting representation $h_{i,k}$ combines aligned Sentinel-2 and weather history with soil, terrain, field, crop, and country context. All checkpoints are fixed before any audit outcome is accessed.

Yield varies across country and crop domains, so we normalize the regression target using fitting statistics. At each checkpoint, an ExtraTrees model and a histogram gradient boosting model predict the normalized yield, and their mixture is mapped back to physical units:

$$\hat{y}_{i,k} = \mu_{d_i} + \sigma_{d_i} \left[(1 - \lambda_k) f_k^{\mathrm{ET}}(h_{i,k}) + \lambda_k f_k^{\mathrm{HGB}}(h_{i,k})\right], \tag{6}$$

where λ_k is selected on a separate farm disjoint selection partition and is then frozen. The low yield threshold for domain d is the 10th percentile of the model fitting yield distribution:

$$\tau_d = Q_{0.20}\left(\{y_i : i \in D_{\text{fit}},\ d_i = d\}\right), \tag{7}$$

and the low yield evidence is

$$s_{i,k} = \frac{\tau_{d_i} - \hat{y}_{i,k}}{\sigma_{d_i}}. \tag{8}$$

Larger scores indicate stronger evidence of low yield, while normalization places fields with different yield scales on a comparable risk scale.

4.2 Sequential Risk Calibration

Let $\mathcal{N}_k = \{s_{j,k} : j \in D_{\text{cal}},\, y_j > \tau_{d_j}\}$ be the null score library at checkpoint k, formed from calibration fields whose final yields exceed their domain thresholds. We convert the current score into a one sided empirical p-value,

$$p_{i,k} = \frac{1 + \sum_{s \in \mathcal{N}_k} \mathbb{1}[s \geq s_{i,k}]}{1 + |\mathcal{N}_k|}. \tag{9}$$

We pre-specify a total field season false watch budget $\alpha = 0.05$ and allocate it across the three checkpoints using fixed error spending weights

$$(w_1, w_2, w_3) = (0.2, 0.3, 0.5), \qquad (\alpha_1, \alpha_2, \alpha_3) = (0.010, 0.015, 0.025). \tag{10}$$

These values are frozen before audit access. We define

$$W_{i,k} = \mathbb{1}[p_{i,k} \leq \alpha_k], \qquad A_{i,k} = W_{i,k-1} W_{i,k}, \quad k \in \{2, 3\}. \tag{11}$$

A single crossing raises a watch, while two consecutive crossings trigger the more conservative action tier. The smallest attainable p-value at checkpoint k is $p_{\min,k} = 1/(1 + |\mathcal{N}_k|)$. The frozen protocol pools the non-low-yield fields from the early-stop and conformal roles, giving $|\mathcal{N}_k| \approx 217$ normal yield reference scores and $p_{\min,k} \approx 0.00459$. Since $\alpha_1 = 0.010 > 0.00459$, Look 1 is triggerable in principle; the later budgets are larger and also triggerable. If $|\mathcal{N}_k| < 99$, then $p_{\min,k} > 0.010$, and Look 1 could never be triggered, so we retain at least 217 normal reference fields in the pooled null library.

Proposition 1 (Field season false watch control)**.** *If a non low yield field is exchangeable with the corresponding null calibration scores at each checkpoint, then*

$$\Pr\left(\exists k : W_{i,k} = 1 \mid z_i = 0\right) \leq \sum_{k=1}^{3} \alpha_k = 0.05. \tag{12}$$

This follows from the super uniformity of the empirical p-values and the union bound, and does not require independence between checkpoints. Since every persistent action is contained in a watch event, the same upper bound also applies to the probability of any false action. All quantities above define the frozen protocol and are fixed before audit.

4.3 Decision Consistent Spatial Refinement

At the final checkpoint, a spatial model predicts within field yield anomalies from the latest available Sentinel-2 observation and environmental variables. Let $\hat{\delta}_{i,p}$ denote the predicted anomaly at pixel p. We construct the $10\,\text{m}$ yield map as

$$\hat{y}_{i,p}^{\text{spatial}} = \hat{y}_{i,3} + \hat{\delta}_{i,p} - \frac{1}{|P_i|} \sum_{q \in P_i} \hat{\delta}_{i,q}. \tag{13}$$

Centering the residuals gives

$$\frac{1}{|P_i|} \sum_{p \in P_i} \hat{y}_{i,p}^{\text{spatial}} = \hat{y}_{i,3}. \tag{14}$$

The spatial branch localizes within field variation while preserving the field level prediction and its associated decision exactly. It does not enter the calibration, watch, or action rules, and is used only as scouting evidence at the final checkpoint.

Table 1: Final-look prediction on the 316 challenge fields. PhysCrop-Risk achieves the lowest RMSE and the highest AUROC and AUPRC. Bold denotes the best result in each column.

Method	RMSE ↓	Pooled R^2 ↑	AUROC ↑	AUPRC ↑	Macro R^2 ↑	Least-favorable R^2 ↑
Country–crop climatology	1.4637	0.7270	0.5524	0.2298	-0.0553	-0.2216
NDVI + NDMI + rain Ridge	1.3254	0.7762	0.7409	0.3783	0.0948	-0.8275
AgriFM encoder + context	1.3268	0.7757	0.6622	0.2913	0.0302	-0.3689
Full-feature Ridge	1.1845	0.8212	0.7352	0.4073	0.2515	-0.5310
MLP	1.2058	0.8147	0.7788	0.4535	0.2947	0.0600
ConvLSTM + context	1.1443	0.8332	0.7688	0.4998	0.2703	-0.2134
Temporal Transformer	1.1493	0.8317	0.7841	0.4906	0.3758	**0.1992**
Random Forest	1.0702	0.8540	0.7950	0.5718	0.3807	0.0695
Phenology-aligned ExtraTrees	1.0459	0.8606	0.8110	0.5717	0.4362	0.1388
PhysCrop-Risk	**1.0361**	**0.8632**	**0.8112**	**0.5735**	**0.4477**	0.1490

5 Experiments

5.1 Experimental Protocol

We evaluate PhysCrop-Risk on the flexible release of YieldSAT (Miranda et al., 2026), which provides irregular Sentinel-2 observations, daily weather, soil and terrain variables, field metadata, final yield, and 10 m harvester yield maps across Argentina, Brazil, Germany, and Uruguay, with corn, rapeseed, soybean, and wheat over nine seasons (2016–2024). After filtering, 2,171 fields remain. We use 1,849 fields for development and reserve 322 for a held-out audit. The development set contains a 1,533-field reference pool and a 316-field challenge set. The reference pool is partitioned by farm identity into 1,254 fields for fitting, 154 for selection, and 125 for calibration, with no farm overlap. The challenge set is used for model comparison, ablations, and spatial evaluation. All preprocessing, checkpoints, predictors, thresholds, calibration, alert policy, and evaluation code were fixed before audit access. Yield prediction uses RMSE and R^2; low-yield ranking uses AUROC and AUPRC; we also report macro and least-favorable-group R^2. Alerts are measured once per field season with FPR, sensitivity, and precision. Prediction comparisons use paired field bootstraps stratified by country and crop, alert intervals use exact Clopper–Pearson intervals, and spatial comparisons treat fields as independent units.

5.2 Prediction and Progress Alignment

Field-level prediction. Table 1 compares all methods on the same 316 challenge fields at the final GDD checkpoint. PhysCrop-Risk achieves the lowest RMSE ($1.0361\ \mathrm{t\,ha^{-1}}$), highest pooled R^2 (0.8632), AUROC 0.8112, AUPRC 0.5735, and macro R^2 0.4477. Its gradient boosting mixture improves ExtraTrees from RMSE 1.0459 to 1.0361 (paired 95% CI $[-0.0163, -0.0038]$), making it the strongest field-level evidence source evaluated.

Progress alignment. Table 2 varies only the temporal coordinate while fixing the causal boundary, feature dimension, and predictor family. Absolute GDD gives the best point estimates (RMSE 1.0361, AUROC 0.8112, AUPRC 0.5735, macro R^2 0.4477), improving over calendar alignment (paired CI $[-0.1328, -0.0001]$), while differences from acquisition rank and normalized season fraction include zero. Modality ablation confirms complementary weather, satellite, and domain-context contributions: removing weather, using Sentinel-2 alone, and removing country/crop context raise RMSE to 1.0936, 1.1419, and 1.3411, respectively (all paired intervals exclude zero).

5.3 Seasonal Evidence and Calibrated Alerts

Evidence strengthens with crop progress. Figure 3 tracks performance across the three GDD checkpoints. Because the number of eligible audit fields increases from 300 at Look 1 to 322 at Look 3, we first evaluate the 300 fields observed at every checkpoint. On this common cohort, RMSE decreases from 1.2407 at Look 1 to 1.0887 at Look 3. Over the same progression, AUROC increases from 0.7370 to 0.8272, and AUPRC increases from 0.4412 to 0.5942. The improvement

Table 2: Temporal-alignment and modality ablations on the 316 challenge fields. All variants use identical causal prefixes and the same predictor family. Intervals in (a) report $\mathrm{RMSE}_{\mathrm{GDD}} - \mathrm{RMSE}_{\mathrm{alternative}}$; intervals in (b) report $\mathrm{RMSE}_{\mathrm{full}} - \mathrm{RMSE}_{\mathrm{ablation}}$. Negative values favor the proposed configuration.

(a) Temporal alignment

Metric	Calendar day	Acquisition rank	Normalized season fraction	**Absolute GDD**
RMSE ↓	1.0961	1.0579	1.0553	**1.0361**
NRMSE ↓	0.3913	0.3776	0.3767	**0.3698**
Skill ↑	0.2511	0.2772	0.2790	**0.2921**
AUROC ↑	0.8045	0.7992	0.8019	**0.8112**
AUPRC ↑	0.5571	0.5361	0.5712	**0.5735**
Macro R^2 ↑	0.4196	0.4223	0.4397	**0.4477**
Paired RMSE 95% CI	$[-0.1328, -0.0001]$	$[-0.0582, 0.0125]$	$[-0.0482, 0.0093]$	–

(b) Input modalities

Representation	RMSE ↓	Paired RMSE 95% CI
Full representation	**1.0361**	–
Without weather	1.0936	$[-0.0876, -0.0324]$
Sentinel-2 only	1.1419	$[-0.1489, -0.0618]$
Without country and crop context	1.3411	$[-0.4246, -0.1895]$

Table 3: **PhysCrop-Risk operating tiers.** Frozen Protocol Results are pre-registered and fixed before audit; Post-hoc Analysis is exploratory, calibration-only, and lacks the finite-sample guarantee of Proposition 1. Intervals are in the appendix; TP/FP denote true/false field-season alerts.

	Watch				Persistent action			
Protocol	TP/FP	FPR	Sens.	Prec.	TP/FP	FPR	Sens.	Prec.
Frozen Protocol Results	25/13	0.0524	0.3378	0.6579	8/5	0.0202	0.1081	0.6154
Post-hoc Analysis	23/12	0.0484	0.3108	0.6571	8/4	0.0161	0.1081	0.6667

therefore follows the accumulation of causal observations rather than a change in the evaluated cohort.

The latest satellite acquisition precedes harvest by a median of 107, 86, and 62.5 days at Looks 1, 2, and 3. On the 322 audit fields at Look 3, PhysCrop-Risk reaches an RMSE of $1.0582\ \mathrm{t\,ha^{-1}}$, an AUROC of 0.8331, and an AUPRC of 0.6245, with RMSE 0.390 lower than climatology.

Frozen protocol alert performance. Table 3 reports the pre-registered operating tiers. The frozen watch tier yields 25 true and 13 false field-season alerts (FPR 5.24%, sensitivity 33.78%, precision 65.79%); its exact 95% Clopper–Pearson interval includes 0.05. Proposition 1 bounds the probability of a false watch, not the realized point estimate. The frozen persistent-action tier yields 8 true and 5 false alerts (FPR 2.02%, sensitivity 10.81%, precision 61.54%), with upper confidence bound below 5%.

Post-hoc calibration-only reference analysis. For exploratory comparison, we additionally report a post-hoc analysis that uses only the 125 designated calibration fields as the null-score library. It yields $23/12$ watch alerts, with an observed FPR of 4.84%, sensitivity of 31.08%, and precision of 65.71%. The persistent-action tier yields $8/4$ alerts, with an observed FPR of 1.61%, sensitivity of 10.81%, and precision of 66.67%. These numbers illustrate the potential operating range of the two tiers, but they do not inherit the finite-sample guarantee of Proposition 1 and are not used for any statistical-validity claim.

Post-hoc analysis and action tier. For exploratory comparison, a post-hoc calibration-only analysis yields watch $23/12$ (FPR 4.84%, sensitivity 31.08%, precision 65.71%) and action $8/4$ (FPR 1.61%, sensitivity 10.81%, precision 66.67%). These numbers do not inherit the finite-sample guarantee and are not used for statistical-validity claims. The action tier's sensitivity of 10.81% reflects a deliberate

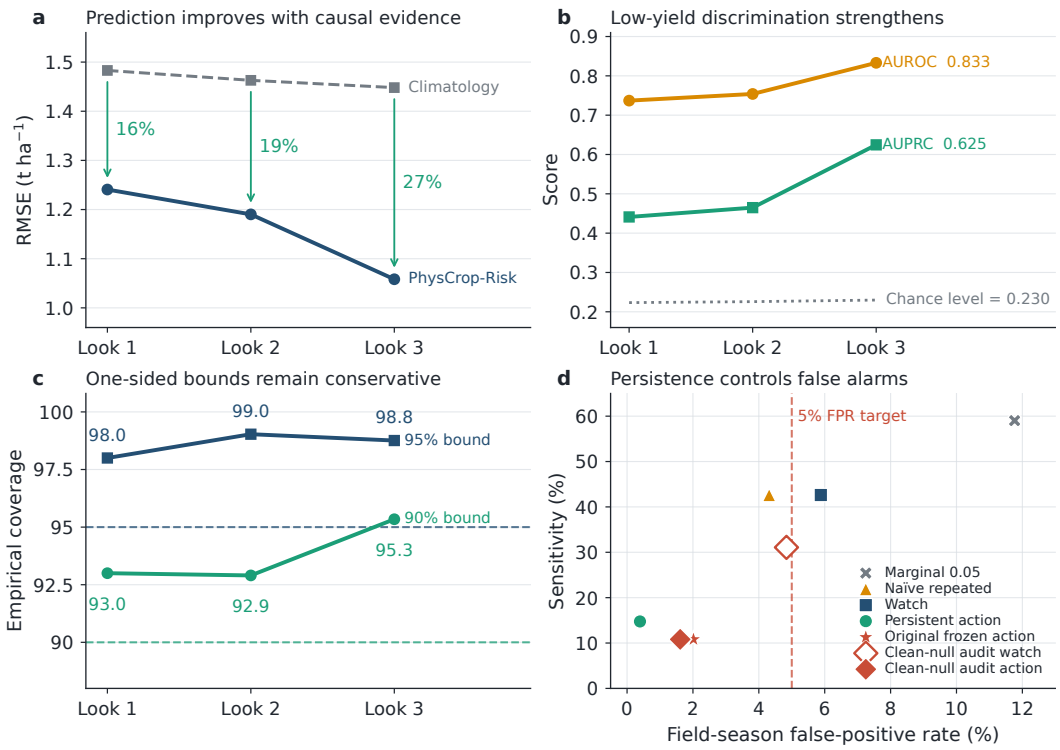

Figure 3: **Prediction and warning evidence across GDD looks.** (a) Audit RMSE for PhysCrop-Risk vs. country-crop climatology; arrows show relative reductions. (b) AUROC/AUPRC for low-yield ranking; final-look prevalence for reference. (c) Empirical coverage of 90%/95% upper yield bounds; dashed lines denote nominal coverage. (d) Sensitivity vs. field-season FPR. Gray/orange/blue/green markers: four policies on 316 challenge fields; red star: frozen action point on 322 audit fields; open/filled diamonds: post-hoc calibration-only watch/action points. Dashed line: 5% target.

trade-off: two consecutive crossings remove transient false alarms and trade recall for precision. The watch tier filters out about 95% of normal fields, while the action tier is more selective (FPR 2.02% under the frozen protocol). Although it identifies only 8 of 74 low-yield fields, its precision is 61.54%, making it a high-confidence confirmation step while the watch tier serves as broader screening.

5.4 Generalization Across Seasons and Farms

We evaluate temporal transfer by holding out each growing season in turn and retraining the predictor on the remaining years. Across the nine held-out seasons, PhysCrop-Risk obtains an RMSE of $1.3251\ \mathrm{t\,ha^{-1}}$ and $R^2 = 0.7790$, compared with a climatology RMSE of $1.6390\ \mathrm{t\,ha^{-1}}$. The paired RMSE difference interval is $[-0.3567, -0.2712]$, showing a consistent improvement across seasons.

We further evaluate generalization across farms using five farm-disjoint folds. The resulting out-of-fold predictions cover $1,848$ fields from 81 farms and achieve an RMSE of $1.2224\ \mathrm{t\,ha^{-1}}$, $R^2 = 0.8120$, AUROC 0.8116, and AUPRC 0.5435. Strong performance under both season-held-out and farm-disjoint evaluation shows that the learned signal transfers beyond the seasons and farms used for fitting.

5.5 Spatial Evidence

Figure 4 evaluates the mean preserving spatial branch on the 316 challenge fields. PhysCrop-Risk achieves a mean within field RMSE of $1.5234\ \mathrm{t\,ha^{-1}}$, improving on the uniform field pre-

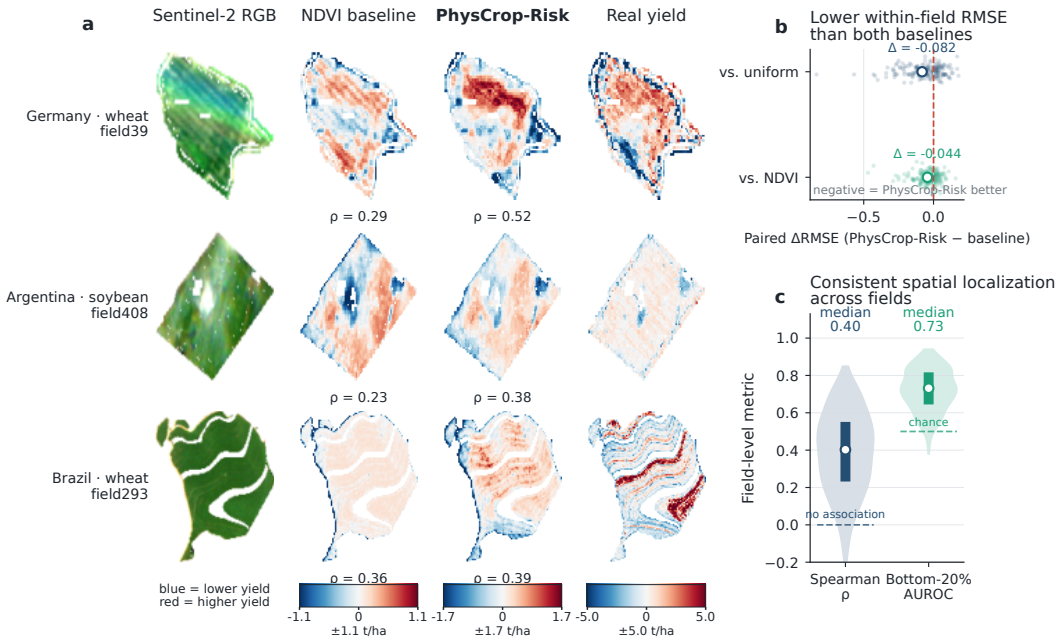

Figure 4: **Spatial evidence at the final GDD look.** (a) Representative fields showing the latest RGB observation, observed yield anomaly, PhysCrop-Risk anomaly, and NDVI baseline. All anomaly maps share the same scale in t ha^{-1}; row labels report the PhysCrop-Risk Spearman correlation. (b) Paired field-level RMSE differences relative to the uniform and NDVI baselines; negative values favor PhysCrop-Risk. (c) Field-wise distributions of Spearman correlation and AUROC for localizing the observed bottom-20% yield pixels; annotations report medians.

diction (1.6053) and an NDVI residual baseline (1.5671), with paired RMSE difference intervals $[-0.0980, -0.0663]$ and $[-0.0528, -0.0348]$, respectively. The predicted maps also recover within field ordering: the median Spearman correlation is 0.4024, and the median AUROC for localizing the bottom 20% yield pixels is 0.7316. The spatial branch adds local structure while preserving the field level prediction used for calibration.

Limitations and exchangeability. Pooling four countries and four crops may violate strict exchangeability, and the pooled null may yield slightly elevated false positive rates in some regions, such as Germany; we therefore do not claim exact regional calibration. Nevertheless, farm-disjoint evaluation over 1,848 fields from 81 farms achieves RMSE 1.2224, R^2 0.8120, AUROC 0.8116, and AUPRC 0.5435, while season-held-out evaluation also improves over climatology. Per-domain calibration would reduce the null library and raise $p_{\min,k}$, motivating our use of pooled domain-normalized scores despite this limitation. Finally, unreached GDD checkpoints are skipped rather than extrapolated, which may reduce later-look coverage in atypical seasons.

6 Conclusion

We presented PhysCrop-Risk, a progress aligned framework for early crop monitoring from partial multimodal observations. GDD checkpoints define causal observation boundaries and decision times, and sequence level calibration controls the field season false watch budget. Across 2,171 fields spanning four countries, four crops, and nine growing seasons, PhysCrop-Risk provides strong early yield prediction and low yield discrimination, with the final checkpoint occurring a median of 62.5 days before harvest. It generalizes across held out seasons and farms, and its mean preserving spatial branch adds within field localization without changing the field level decision.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, literature search, and reference formatting. They were also used to discuss the presentation and interpretation of experimental results already produced by the authors. All cited literature was independently checked against publication records, and all AI-assisted text was reviewed and revised by the authors. The authors verified the reported methods, mathematical statements, experimental results, and references, and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the dataset construction, data partitions, causal evaluation protocol, progress alignment procedure, predictive models, calibration rule, and evaluation metrics in the main paper. The appendix provides additional implementation details, hyperparameters, preprocessing specifications, and the proof of Proposition 1. All quantities used for the held out audit, including progress checkpoints, domain thresholds, predictors, calibration settings, and alert rules, were fixed before accessing the audit outcomes. Code and configuration files needed to reproduce the experiments will be provided in the anonymized repository`https://anonymous.4open.science/r/physcrop-9A96`.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024a.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1616–1630. PMLR, 2024b.

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. OmniSat: Self-supervised modality fusion for earth observation. In *Computer Vision – ECCV 2024*, pp. 409–427, 2024.

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19530–19540, 2025.

Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27788–27799, 2024.

Raymond Bonhomme. Bases and limits to using 'degree.day' units. *European Journal of Agronomy*, 13(1):1–10, 2000. doi: 10.1016/S1161-0301(00)00058-7.

Baiting Chen, Zhimei Ren, and Lu Cheng. Conformalized time series with semantic features. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9439–9457. PMLR, 2024.

Minh Kha Do, Kang Han, Phu Lai, Khoa T. Phan, and Wei Xiang. RobSense: A robust multimodal foundation model for remote sensing with static, temporal, and incomplete data adaptability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7427–7436, 2025.

Keke Duan, Anton Vrieling, Michael Schlund, Uday Bhaskar Nidumolu, Christina Ratcliff, Simon Collings, and Andrew Nelson. Detection and attribution of cereal yield losses using Sentinel-2 and weather data: A case study in south australia. *ISPRS Journal of Photogrammetry and Remote Sensing*, 213:33–52, 2024. doi: 10.1016/j.isprsjprs.2024.05.021.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, 2024.

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*, pp. 34818–34831, 2026.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020.

Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. MAESTRO: Masked autoencoders for multimodal, multitemporal, and multispectral earth observation data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 212–224, 2026.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983. doi: 10.1093/biomet/70.3.659.

Wenyuan Li, Shunlin Liang, Keyan Chen, Yongzhe Chen, Han Ma, Jianglei Xu, Yichuan Ma, Yuxiang Zhang, Shikang Guan, Husheng Fang, and Zhenwei Shi. AgriFM: A multi-source temporal remote sensing foundation model for agriculture mapping. *Remote Sensing of Environment*, 334: 115234, 2026. doi: 10.1016/j.rse.2026.115234.

Yi Liu, Alexander Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31344–31382. PMLR, 2024.

Yu Luo, Xiaogang Zhu, Shan Zeng, Wei Xiang, Thomas Francis Bishop, Zhiyong Wang, and Kun Hu. PhenoYieldNet: Learning crop-aware phenological responses for multi-crop yield prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15870–15879, 2026.

Gregory S. McMaster and W. W. Wilhelm. Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology*, 87(4):291–300, 1997. doi: 10.1016/S0168-1923(97)00027-0.

Francisco Mena, Deepak Pathak, Hiba Najjar, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Miro Miranda, Jayanth Siddamsetty, Marlon Nuske, Marcela Charfuelan, Diego Arenas, Michaela Vollmer, and Andreas Dengel. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. *Remote Sensing of Environment*, 318:114547, 2025. doi: 10.1016/j.rse.2024.114547.

Miro Miranda, Deepak Pathak, Patrick Helber, Benjamin Bischke, Hiba Najjar, Francisco Mena, Cristhian Sanchez, Akshay Pai, Diego Arenas, Matias Valdenegro-Toro, Marcela Charfuelan, Marlon Nuske, and Andreas Dengel. YieldSAT: A multimodal benchmark dataset for high-resolution crop yield prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22920–22930, 2026.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *Computer Vision – ECCV 2024*, pp. 164–182, 2024.

Aleksandr Podkopaev, Dong Xu, and Kuang-Chih Lee. Adaptive conformal inference by betting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40886–40907. PMLR, 2024.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan Rao Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Jinjin Tian and Aaditya Ramdas. Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30(4):976–993, 2021. doi: 10.1177/0962280220983381.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60280–60300. PMLR, 2025.

Lars van der Laan and Ahmed M. Alaa. Self-calibrating conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Janaksinh Ven, Charu Sharma, and Azeemuddin Syed. SegMango: Early deep mango yield prediction based on flower segmentation and weather data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4984–4993, 2026.

Guilong Xiao, Xueyou Zhang, Quandi Niu, Xingang Li, Xuecao Li, Liheng Zhong, and Jianxi Huang. Winter wheat yield estimation at the field scale using Sentinel-2 data and deep learning. *Computers and Electronics in Agriculture*, 216:108555, 2024. doi: 10.1016/j.compag.2023.108555.

Xingguo Xiong, Renhai Zhong, Qiyu Tian, Jingfeng Huang, Linchao Zhu, Yi Yang, and Tao Lin. Daily DeepCropNet: A hierarchical deep learning approach with daily time series of vegetation indices and climatic variables for corn yield estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209:249–264, 2024. doi: 10.1016/j.isprsjprs.2024.02.008.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55076–55099. PMLR, 2024.

Hossein Zare, Tobias K. D. Weber, Joachim Ingwersen, Wolfgang Nowak, Sebastian Gayler, and Thilo Streck. Within-season crop yield prediction by a multi-model ensemble with integrated data assimilation. *Field Crops Research*, 308:109293, 2024. doi: 10.1016/j.fcr.2024.109293.

Qu Zhou, Kaiyu Guan, Sheng Wang, James Hipple, and Zhangliang Chen. From satellite-based phenological metrics to crop planting dates: Deriving field-level planting dates for corn and soybean in the u.s. midwest. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216:259–273, 2024. doi: 10.1016/j.isprsjprs.2024.07.031.

APPENDIX CONTENTS

1. A Proof of Proposition 1
2. B Calibration Details and Triggerability
3. C Implementation Details
4. D Additional Experimental Results
5. E Compute and Runtime
6. F Reproducibility Checklist

A PROOF OF PROPOSITION 1

Fix a truly non-low-yield field i ($z_i = 0$) and a checkpoint k. The null library $\mathcal{N}_k$ contains the risk scores $s_{j,k}$ of held-out reference fields with $z_j = 0$, and the evaluation score $s_{i,k}$ is computed from the same frozen predictor and the same normalisation. Under the protocol's exchangeability assumption, the reference and evaluation scores are exchangeable, so the rank of $s_{i,k}$ among $\{s_{i,k}\} \cup \mathcal{N}_k$ is uniform. With the finite-sample correction used in the paper,

$$p_{i,k} = \frac{1 + \sum_{j \in \mathcal{N}_k} \mathbb{1}[s_{j,k} \geq s_{i,k}]}{1 + |\mathcal{N}_k|},$$

it follows that for every $t \in [0, 1]$,

$$\Pr\left(p_{i,k} \leq t \mid z_i = 0\right) \leq t,$$

that is, each $p_{i,k}$ is super-uniform. Applying the union bound over the three checkpoints, and writing $W_{i,k} = \mathbb{1}[p_{i,k} \leq \alpha_k]$, gives

$$\Pr\left(\exists k : W_{i,k} = 1 \,\middle|\, z_i = 0\right) = \Pr\left(\exists k : p_{i,k} \leq \alpha_k \,\middle|\, z_i = 0\right) \leq \sum_{k=1}^{3} \alpha_k = 0.05.$$

The argument requires neither independence across checkpoints nor any assumption on the joint distribution beyond exchangeability; it is a union bound over the three p-values.

The weights α_k are not chosen post hoc. Our development experiments (Appendix D) compare five pre-registered policies on the 316-field development challenge set, without reading any audit outcome: a naive rule, a per-checkpoint $\alpha = 0.05$ rule, an equal Bonferroni split ($\alpha_k = 1/60$), the spending rule $(0.010, 0.015, 0.025)$ used here, and the same spending rule with a two-hit persistence requirement. The spending rule allocates more budget to later checkpoints, which carry more information, and was selected together with persistence as the operating policy; the total budget remains 0.05, so Proposition 1 applies unchanged. Persistence can only remove alerts: $A_{i,k} = W_{i,k-1} W_{i,k} \leq W_{i,k}$, hence $\Pr(\exists k : A_{i,k} = 1 \mid z_i = 0) \leq \Pr(\exists k : W_{i,k} = 1 \mid z_i = 0) \leq 0.05$.

The guarantee is stated for the pre-registered protocol, whose null library pools the non-low-yield early-stop and conformal reference fields. The clean-null analysis in Appendix B replaces this library with conformal-role fields only; it is an exploratory, post-hoc comparison motivated by the fact that the early-stop partition participates in mixture selection, and it does not inherit the formal finite-sample statement above.

B CALIBRATION DETAILS AND TRIGGERABILITY

Roles and null libraries. The development reference pool defines three farm-disjoint roles: a 1,254-field *model-fit* partition used to fit the predictors, a 154-field *early-stop* partition used only to select the mixture weights λ_k, and a 125-field *conformal* partition used only for one-sided calibration and for the alert null. A field must have at least three usable acquisitions before a checkpoint to be eligible; the eligible conformal support is therefore 118, 125, and 125 fields at Looks 1–3 instead of the nominal 125. The pre-registered (frozen) protocol builds its alert null by pooling the non-low-yield fields of the early-stop and conformal roles, which yields 216, 217, and 217 reference scores. The post-hoc clean-null analysis rebuilds the null from the conformal role alone, which

Table 4: Null-library size and the minimum attainable p-value at each checkpoint. A checkpoint is triggerable by any watch only if $p_{\min,k} \leq \alpha_k$, i.e. $|\mathcal{N}_k| \geq \lceil 1/\alpha_k \rceil - 1$.

Protocol	Checkpoint	$\lvert\mathcal{N}_k\rvert$	$p_{\min,k} = \frac{1}{1+\lvert\mathcal{N}_k\rvert}$	α_k	Triggerable
Frozen (early-stop + conformal non-low)	Look 1	216	0.00461	0.010	yes
	Look 2	217	0.00459	0.015	yes
	Look 3	217	0.00459	0.025	yes
Clean null (conformal non-low only)	Look 1	96	0.01031	0.010	**no**
	Look 2	97	0.01020	0.015	yes
	Look 3	97	0.01020	0.025	yes

yields 96, 97, and 97 non-low-yield reference scores. The value 125 therefore denotes the size of the calibration partition, not the size of the null library used by either protocol. Table 4 summarises the resulting libraries and minimum attainable p-values.

Two consequences deserve emphasis. First, the frozen protocol can trigger at Look 1 because its pooled library has more than $\lceil 1/0.010 \rceil - 1 = 99$ non-low-yield fields. Second, under the clean null the conformal-only library has 96 non-low-yield fields at Look 1, so $p_{\min,1} = 1/97 = 0.01031 > \alpha_1 = 0.010$ and *Look 1 cannot fire any watch*. All clean-null watch and action events therefore originate from Looks 2–3, which is visible in the stage-wise table below. We report this explicitly rather than treating the clean null as a drop-in replacement; it is a stricter, exploratory reference analysis.

If calibration were performed separately for every country–crop cell, several domains would contain fewer than 99 non-low-yield reference fields and Look 1 would be untriggerable there. We therefore use pooled, domain-normalised null scores and list the absence of per-domain null libraries as a limitation.

Stage-wise empirical behaviour. Table 5 reports per-checkpoint watch behaviour. The frozen protocol accumulates evidence from Look 1; the clean-null analysis starts at Look 2 by construction. Importantly, the table counts a field at *every* checkpoint at which it is flagged, whereas Table 3 counts a field-season as positive if it is flagged at least once across the season. A field can be flagged at more than one checkpoint, so the stage-wise true/false positives (e.g. $5{+}10{+}22$ and $3{+}8{+}9$ for the frozen protocol) do not sum to the season-level counts in Table 3; they are two different aggregations of the same decisions.

Table 5: Stage-wise watch behaviour on the 322-field audit. Fractions are computed among the fields eligible at that checkpoint (n shown in the last column); TP/FP count low-yield and non-low-yield fields flagged at that checkpoint.

Protocol	Checkpoint	TP	FP	FPR	Sens.	Prec.	n
Frozen	Look 1	5	3	0.0129	0.0746	0.625	300
Frozen	Look 2	10	8	0.0333	0.1429	0.556	310
Frozen	Look 3	22	9	0.0363	0.2973	0.710	322
Clean null	Look 1	0	0	0.0000	0.0000	–	300
Clean null	Look 2	10	7	0.0292	0.1429	0.588	310
Clean null	Look 3	21	9	0.0363	0.2838	0.700	322

Exact intervals. Table 6 lists two-sided exact Clopper–Pearson intervals for the two protocols in Table 3. For the frozen watch tier the FPR interval $[0.0282, 0.0880]$ contains the nominal 0.05; the corresponding one-sided upper bound, which is the quantity used by the pre-registered gate for the action tier, is below its audit threshold. The post-hoc intervals are the same exact intervals recomputed on the clean-null decisions; farm-clustered bootstrap intervals are reported in the main text.

Table 6: Two-sided exact 95% Clopper–Pearson intervals for the decisions in Table 3.

Protocol	Tier	FPR [95% CI]	Sens. [95% CI]	Prec. [95% CI]
Frozen	Watch	$[0.0282, 0.0880]$	$[0.2319, 0.4572]$	$[0.4865, 0.8037]$
Frozen	Persistent action	$[0.0066, 0.0464]$	$[0.0478, 0.2020]$	$[0.3158, 0.8614]$
Post-hoc	Watch	$[0.0252, 0.0830]$	$[0.2083, 0.4290]$	$[0.4779, 0.8087]$
Post-hoc	Persistent action	$[0.0044, 0.0408]$	$[0.0478, 0.2020]$	$[0.3489, 0.9008]$

C Implementation Details

Inputs and preprocessing. For every field and acquisition we form twelve Sentinel-2 surface-reflectance bands (B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, B12), four spectral indices, the valid-pixel fraction, cumulative precipitation, and cumulative growing-degree days. The indices are $\mathrm{NDVI} = (B08-B04)/(B08+B04)$, $\mathrm{NDMI} = (B08-B11)/(B08+B11)$, $\mathrm{NDRE} = (B07-B05)/(B07+B05)$, and $\mathrm{EVI} = 2.5(B08-B04)/(B08+6B04-7.5B02+1)$. Band and index values are averaged over pixels with SCL classes $\{4, 5\}$ inside the field boundary. Daily weather is converted from Kelvin to Celsius and from metres to millimetres when required (values above 100 are treated as Kelvin; precipitation is clipped at zero). Daily thermal accumulation is $\max(0, \min(25, T_{\mathrm{mean}} - 5))$ degree-days, accumulated from sowing up to the information cutoff. Two fields in the 2,173-field archive have fewer than four usable acquisitions and are removed at dataset construction (2,171 fields remain). At checkpoint level a field additionally requires at least three usable acquisitions on or before the cutoff; one structurally ineligible audit field is excluded before the freeze. Missing feature values are imputed with model-fit medians; a channel observed only once is broadcast to all thermal-time positions.

Thermal-time checkpoints and interpolation grid. Checkpoint k for crop c is $L_{c,k} = \rho_k \,\mathrm{median}_{i \in \mathcal{D}_{\mathrm{fit}}, c_i = c}(\max_t g_{it})$ with the pre-registered fractions $(\rho_1, \rho_2, \rho_3) = (0.35, 0.50, 0.65)$; the resulting absolute-GDD thresholds (e.g. 926/1323/1720 for corn and 951/1358/1766 for soybean) are written to the released milestone file. Only observations with $g_{it} \le L_{c,k}$ enter checkpoint k. Each dynamic channel is linearly interpolated in thermal time onto the six equally spaced positions $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\} \times L_{c,k}$. This is interpolation of the available trajectory, not binning or averaging: the six positions sample a piecewise-linear reconstruction of all acquisitions before the cutoff, and the cumulative channels retain the timing information that a pure snapshot would lose. The alignment-coordinate study in Appendix D compares this choice with calendar day-of-season, acquisition rank, and normalised observed-season fraction under an identical causal prefix and feature budget (Table 7); absolute GDD achieves the best RMSE, and the paired difference against calendar time excludes zero.

Feature vector. The dynamic block contributes $6 \times 18 = 108$ values (the twelve bands, four indices, valid fraction and cumulative precipitation; cumulative GDD is the coordinate). The static block contributes 15 values (eight soil properties, five terrain variables, field area and circularity), and context contributes eight one-hot indicators (four countries, four crops), for 131 inputs in total.

Predictors and mixture. Each checkpoint fits two regressors on the model-fit partition: an ExtraTrees regressor (600 trees, minimum leaf size 3, 0.7 feature subsampling, no bootstrap, `random_state`=2201) and a histogram gradient-boosting regressor (300 iterations, minimum leaf size 15, $\ell_2 = 1.0$, `random_state`=5501). Targets are z-normalised within country–crop cells using model-fit means and standard deviations, with the standard deviation floored at $0.25\,\mathrm{t\,ha^{-1}}$ to avoid unstable normalisation for homogeneous groups. The mixture weight $\lambda_k \in \{0, 0.1, \ldots, 1\}$ minimises RMSE on the farm-disjoint early-stop partition; the selected values are $\lambda = (0, 0, 0.1)$ for Looks 1–3, so the first two checkpoints use ExtraTrees only and the final checkpoint adds a 10% boosting component.

Low-yield definition. The low-yield cutoff is the 20th percentile (q_{20}) of model-fit yields, computed in a pre-declared hierarchy: country–crop cells with at least 30 fields, otherwise the crop distribution when it has at least 50 fields, otherwise the global distribution. Written explicitly, $\tau_d = Q_{0.20}\big(\{y_i : i \in \mathcal{D}_{\mathrm{fit}},\ d_i = d\}\big)$ within the applicable hierarchy level, and the binary end-

point is $z_i = \mathbb{1}[y_i \leq \tau_{d_i}]$. We use q_{20} rather than a lower percentile because the endpoint is defined as training-relative bottom-quintile yield; the resulting thresholds are frozen before the audit and are included in the released reference file.

Calibration, p-values, and alert policy. One-sided split-conformal bounds use global raw residuals at Looks 1 and 3 and hierarchical residuals scaled by the square root of tree dispersion at Look 2. Alert decisions use class-conditional empirical p-values with the $+1$ correction of Appendix A, computed against non-low-yield reference fields only. The alpha budget $(0.010, 0.015, 0.025)$ and the persistence requirement were selected on development data (Appendix D, Table 9) against a naive rule, a flat per-checkpoint 0.05 rule, and an equal Bonferroni split; a single crossing defines a watch and two consecutive crossings define a persistent action. The clean-null analysis repeats the same computation with the null library restricted to conformal-role fields.

Spatial residual model. The spatial branch consumes, per pixel, the twelve bands, four indices, eight soil layers, and five terrain layers (29 features), together with the frozen field-level prediction. A histogram gradient-boosting regressor (300 iterations, minimum leaf size 50, $\ell_2 = 1.0$, `random_state`=7701) is trained on at most 256 randomly sampled pixels per model-fit field to predict deviations from that field's observed mean yield; an NDVI-only ridge on the same grid serves as the baseline. At inference the predicted residuals are centred to zero mean before being added to the field prediction, so the field-level mean is preserved exactly (maximum deviation 1.8×10^{-15} t ha^{-1} across the 316 evaluated fields). Maps are produced at 10 m resolution from the latest valid acquisition no later than $L_{c,3}$; statistical comparisons resample fields, never pixels.

D Additional Experimental Results

Alignment coordinate. Table 7 holds the feature budget, causal prefix, and model family fixed while only the alignment coordinate changes. Absolute GDD gives the best RMSE, normalised RMSE, climatology skill, AUROC, AUPRC, and macro-group R^2.

Table 7: Alignment-coordinate comparison on the 316-field development challenge set.

Alignment	RMSE	NRMSE	Skill	AUROC	AUPRC	Macro R^2	Worst R^2
Calendar day-of-season	1.0961	0.3913	0.2511	0.8045	0.5571	0.4196	0.2788
Acquisition rank	1.0579	0.3776	0.2772	0.7992	0.5361	0.4223	0.1762
Normalised season fraction	1.0553	0.3767	0.2790	0.8019	0.5712	0.4397	0.1965
Absolute GDD	**1.0361**	**0.3698**	**0.2921**	**0.8112**	**0.5735**	**0.4477**	0.1490

Modality and component ablations. Table 8 removes one input block at a time from the frozen pipeline. Weather, static context, spectral indices, and domain indicators all contribute; the single-tree variant and the rank-aligned variant quantify the value of the dual-tree mixture and of thermal-time alignment.

Table 8: Prediction ablations on the 316-field development challenge set.

Variant	RMSE	R^2	AUROC	AUPRC	Macro R^2	Worst R^2
Full model	**1.0361**	**0.8632**	**0.8112**	0.5735	**0.4477**	0.1490
No weather	1.0936	0.8476	0.7835	0.5117	0.3934	0.0462
No static context	1.0496	0.8596	0.8100	0.5698	0.4346	0.1351
No spectral indices	1.0521	0.8589	0.8037	**0.5859**	0.4265	0.0646
Sentinel-2 only	1.1419	0.8339	0.7760	0.4702	0.3505	−0.1279
No domain indicators	1.3411	0.7708	0.7691	0.4188	0.1970	−0.1630
ExtraTrees only	1.0459	0.8606	0.8110	0.5717	0.4362	0.1388
Rank-aligned features	1.0579	0.8574	0.7992	0.5361	0.4223	0.1762

Alpha-allocation sensitivity. Table 9 re-allocates the same total budget of 0.05 across checkpoints under the clean null. Equal Bonferroni spending yields slightly higher sensitivity at a slightly higher

FPR; a front-loaded allocation is worse on both tiers; a flat 0.05 per checkpoint violates the family-wise budget and roughly triples the action FPR. The spending rule used in the paper provides the best precision–reliability trade-off among the compared allocations and was the policy selected on development data.

Table 9: Alpha-allocation sensitivity under the clean null (total budget 0.05 except the flat comparator, which is diagnostic).

Allocation $(\alpha_1, \alpha_2, \alpha_3)$	Watch			Persistent action		
	FPR	Sens.	Prec.	FPR	Sens.	Prec.
Frozen spending $(0.010, 0.015, 0.025)$	**0.0484**	0.3108	**0.6571**	**0.0161**	0.1081	**0.6667**
Equal Bonferroni $(1/60, 1/60, 1/60)$	0.0524	**0.3378**	0.6579	0.0202	**0.1216**	0.6429
Front-loaded $(0.025, 0.015, 0.010)$	0.0605	0.2297	0.5312	0.0242	0.0946	0.5385
Flat 0.05 per checkpoint	0.1452	0.4324	0.4706	0.0726	0.2568	0.5135

Per-group alert behaviour. Table 10 reports the clean-null persistent action per supported country–crop group. Zero FPR is attainable in most groups, while sensitivity is concentrated in the three soybean groups; several small groups contain no action-level true positive, which is consistent with the deliberately conservative action tier and with the narrow worst-group audit failure reported in the main text.

Table 10: Clean-null persistent action by country–crop group (audit set).

Group	n	Low	TP	FP	FPR	Sens.
Argentina corn	29	4	0	0	0.0000	0.0000
Argentina soybean	65	15	1	1	0.0200	0.0667
Argentina wheat	23	5	0	0	0.0000	0.0000
Brazil corn	22	4	0	0	0.0000	0.0000
Brazil soybean	55	21	5	1	0.0294	0.2381
Brazil wheat	14	1	0	0	0.0000	0.0000
Germany rapeseed	16	5	0	0	0.0000	0.0000
Germany wheat	23	5	0	0	0.0000	0.0000
Uruguay soybean	75	14	2	2	0.0328	0.1429

False negatives. Under the clean-null action tier, 66 of 74 low-yield audit fields are not alerted. Their median Look-3 p-value is 0.112 (interquartile range 0.051–0.327), i.e. most misses are fields whose evidence never approached the action threshold rather than borderline decisions. The largest over-predictions among misses are Argentina corn (predicted 5.21 t ha^{-1} above the observed yield), followed by three wheat fields (3.27–4.75 t ha^{-1}). The misses are spread across all nine groups (largest counts: Brazil soybean 16, Argentina soybean 14, Uruguay soybean 12), which is the expected behaviour of a high-precision, low-recall operating tier whose sensitivity is 10.8%; we report the trade-off rather than tuning the tier on the audit set.

Spatial evidence. The spatial branch is evaluated on 316 development-challenge fields and 1,530,144 pixels with field-clustered inference. The mean field RMSE is 1.5234 for the residual map, 1.6053 for a uniform field prediction, and 1.5671 for an NDVI-only residual model; both paired field-bootstrap intervals exclude zero. Across fields, the median within-field Spearman correlation is 0.4024 and the median AUROC for locating the observed bottom-yield quintile is 0.7316. Additional qualitative fields and residual distributions are provided with the released maps.

E Compute and Runtime

All experiments were run on a CPU-only machine (12th Gen Intel Core i5-1240P, 12 cores / 16 threads, PyTorch 2.12.0+cpu, no CUDA device). Table 11 reports measured wall-clock costs for the proposed system. Training times cover one fit of each component on the model-fit partition; the mixture weight is selected by evaluating 11 candidate weights on the farm-disjoint early-stop

partition, which costs a small constant multiple of a single prediction pass and is included in the training column.

Table 11: Measured compute cost of the proposed system on CPU.

Checkpoint	Fit fields	ExtraTrees (s)	HGB (s)	Inference (ms/field)	Model size (MB)
Look 1	1,191	1.61	5.25	–	29.9
Look 2	1,230	1.69	5.31	–	30.9
Look 3	1,254	0.74	1.46	0.43	31.4
Spatial residual model	1,254	–	–	–	1.15

Inference for both tree components at Look 3 averages 0.43 ms per field on the same CPU (batch of 322 audit fields, five repetitions), with negligible conformal and decision overhead; the full three-checkpoint pipeline therefore runs in milliseconds per field-season. The spatial branch is the only component whose cost scales with field area rather than field count, since it predicts per-pixel residuals; it is optional and does not affect the field-level alert decisions. The deep visual baselines (frozen AgriFM encoder, ConvLSTM, temporal transformer) were trained with the released scripts on the same CPU-only machine and are the dominant compute cost of the comparison study; they are not required to reproduce the alerting results.

F Reproducibility Checklist

- **Code.** The full pipeline and the report agent are released as a single anonymous archive: reproduction scripts for the feature cache, Experiments 1–9, the vision baselines, the frozen audit, the post-hoc analyses, and the figure generation, together with the frozen artifacts needed to replay the audit. The report agent that consumes the frozen evidence (rules plus a constrained language model) is included in the same archive and is not required for any number in this paper.
- **Configurations.** All protocol constants (GDD fractions, interpolation grid, alpha spending, persistence, hierarchy thresholds, hyperparameters) are stored in the scripts and in `MASTER_PROTOCOL.md`; the frozen audit additionally writes a manifest with the SHA-256 hash of every input, model, and protocol file.
- **Random seeds.** ExtraTrees `random_state`=2201; histogram gradient boosting `random_state` = 5501 (tabular) and 7701 (spatial); the role split uses a fixed SHA-256 hash salt; field-bootstrap and farm-cluster bootstrap analyses use fixed seeds (999,001 and 2026,0801). Prediction caches are deterministic given the seeds.
- **Data.** We use the public YieldSAT release (Miranda et al., 2026; `arXiv:2604.00940`). The exact archive layout expected by the scripts is documented in the release README; no raw data are redistributed. Frozen feature caches and audit models are included so that the one-shot audit can be replayed without downloading the raw archive.
- **Splits.** Role assignment (model-fit, early-stop, conformal, locked challenge, audit) is deterministic and derived only from farm/field identifiers; the audit field list and its hash are frozen in the manifest before outcome reveal.
- **Evaluation.** `run_yieldsat_final_audit.py` performs the one-shot audit and refuses to run twice; `run_yieldsat_postaudit_p0a.py`, `p0b.py`, and `p1.py` reproduce the post-hoc analyses; `generate_yieldsat_paper_figures.py` regenerates Figures 3–4.
- **Environment.** Python 3.10/3.12 with numpy, scipy, scikit-learn, joblib, tifffile, matplotlib, and PyTorch (CPU build) for reproducibility; the report agent additionally uses a lightweight PDF library.