PHENOBIID: RETHINKING IN-SEASON CROP YIELD PREDICTION FROM A WORLD MODEL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Crop yield is the terminal outcome of a dynamic growth process, where weather drives vegetation evolution and accumulated seasonal development determines harvest productivity. However, during the growing season, only part of this process is observable before harvest, making it challenging to predict yield from incomplete growth information. Existing in-season yield prediction methods mainly aggregate available observations and directly regress final yield, leaving the hidden evolution of vegetation dynamics insufficiently modeled. In this work, we reformulate in-season yield prediction as a predictive world-modeling problem and introduce PhenoBIID, a task-specific weather-conditioned vegetation world model that learns weather-driven remote sensing vegetation state transitions and predicts hidden growth trajectories. Instead of directly mapping partial seasonal observations to yield, PhenoBIID reconstructs the unobserved portion of vegetation development and uses the completed trajectory for yield estimation, explicitly connecting environmental forcing, vegetation dynamics, and crop outcomes. To evaluate this perspective, we introduce CropDynamicsBench, a global benchmark covering maize, rice, soybean, and wheat that jointly evaluates vegetation trajectory forecasting and yield prediction under varying observation availability. In retrospective evaluations under realized weather, PhenoBIID reduces seed-mean yield RMSE by 0.12–4.74% relative to the strongest evaluated direct baseline at each crop–cutoff setting with 10–70% unobserved windows. Our findings highlight the potential of predictive world modeling to connect crop growth dynamics with yield assessment and support agricultural planning. Code, processed benchmark inputs, an executable inference sample, and evaluation artifacts are available at `https://anonymous.4open.science/r/PhenoBIID-5EBB`.

1 INTRODUCTION

Timely crop yield prediction before harvest supports food security assessment and agricultural planning (Anderson et al., 2024; Chen et al., 2026), particularly in the context of climate change and agricultural adaptation (Hultgren et al., 2025). However, crop yield is the terminal outcome of a dynamic growth process rather than a static attribute of the crop state at the time of observation. Environmental conditions continuously shape crop development, biomass accumulation, and grain filling, thereby influencing final yield (Minoli et al., 2022; Lobell et al., 2012). During the growing season, only a partial trajectory of crop development is observed, while subsequent vegetation development remains uncertain. A fundamental question therefore arises: *under given weather conditions, can forecasting the unobserved part of vegetation development improve yield assessment?*

Existing in-season yield prediction methods typically formulate the task as a direct regression problem, where satellite observations, meteorological variables, and historical yield records are aggregated to predict final yield (Schwalbert et al., 2020; Kallenberg et al., 2026). Recent advances in multimodal learning and spatiotemporal modeling have improved the integration of heterogeneous environmental signals. However, these approaches mainly exploit currently available information and treat vegetation observations as input features, while the unobserved evolution between the observation cutoff and harvest remains insufficiently modeled. A vegetation signal that is informative when directly observed may not retain the same value when it must be forecasted, and its utility can vary with crop and observation time (Johnson, 2016; Lu et al., 2026). Therefore, beyond extracting

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.

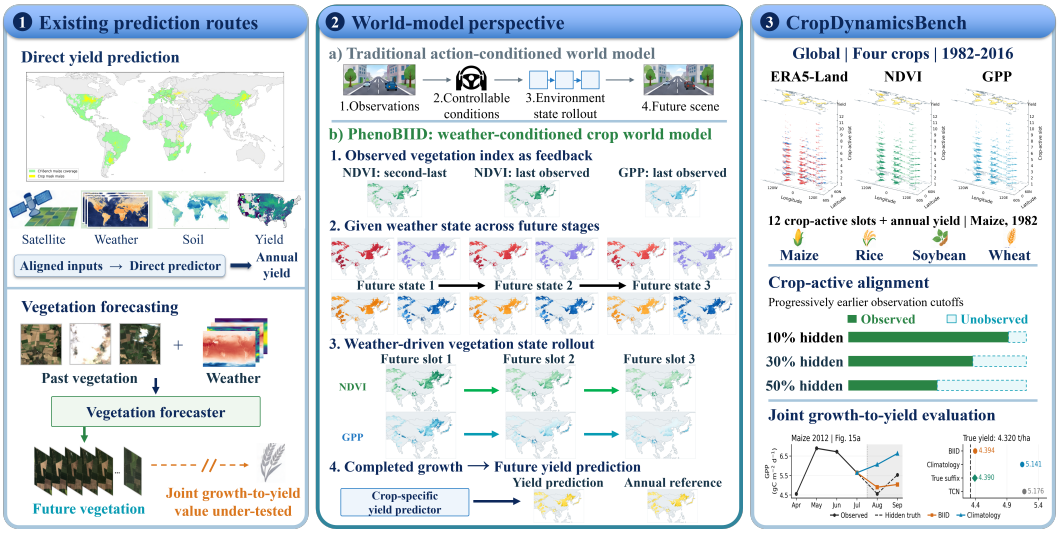

Figure 1: A world-model view of in-season yield prediction: CROPDYNAMICSBENCH and PHENOBIID connect weather-driven growth forecasting to yield evaluation.

information from observations, effective yield prediction requires understanding and modeling the underlying vegetation dynamics.

A natural perspective for addressing this challenge is to view crop development as a weather-conditioned dynamical system. In this view, weather provides external forcing, vegetation states represent the evolving condition of crops, and yield emerges as the final outcome of the growth trajectory. This formulation aligns with the concept of predictive world models, where models learn latent state transitions and forecast future states under changing conditions. Although recent Earth observation world models have demonstrated promising capabilities in environmental state forecasting (Benson et al., 2024; Iele et al., 2026; Luo et al., 2026a), their evaluations mainly focus on trajectory reconstruction accuracy, leaving an important question unanswered: *whether predicted future states can provide meaningful benefits for downstream agricultural prediction tasks.*

Motivated by this perspective, we introduce PhenoBIID, a task-specific weather-conditioned vegetation world model for in-season crop yield prediction. Instead of directly mapping partially observed seasonal information to final yield, PhenoBIID models crop development as a dynamic process driven by environmental conditions, predicts the hidden portion of vegetation evolution, and leverages the completed trajectory for yield estimation. This formulation explicitly connects weather forcing, vegetation dynamics, and crop outcomes, transforming in-season yield prediction from a static regression problem into a predictive modeling problem of hidden growth processes. To systematically evaluate this perspective, we establish CropDynamicsBench, an evaluation platform that jointly assesses vegetation trajectory forecasting and downstream yield prediction under different observation availability conditions. Beyond final yield accuracy, the benchmark enables analysis of whether predicted future vegetation states provide additional predictive value compared with directly using available observations. Retrospective experiments under realized weather show that PhenoBIID improves in-season yield prediction, with predictive vegetation completion providing additional benefits for several crops, especially over longer unobserved windows.

Our contributions are summarized as follows:

- We formulate in-season crop yield prediction from a predictive world-model perspective, where hidden vegetation evolution is explicitly modeled as an intermediate process between weather conditions and final yield.
- We propose PhenoBIID, a task-specific weather-conditioned vegetation world model that learns latent vegetation dynamics, predicts unobserved growth trajectories, and connects trajectory forecasting with downstream yield estimation.
- We introduce CropDynamicsBench, a benchmark for jointly evaluating vegetation trajectory forecasting and yield prediction under different observation availability conditions, enabling systematic study of the value of predictive growth dynamics.

2 Related Work

2.1 In-season Crop Yield Prediction

In-season crop yield prediction aims to estimate final agricultural productivity before harvest by integrating available environmental and historical information. Early approaches mainly relied on statistical models and machine learning methods using historical yield records, climate variables, and remote sensing observations (Schwalbert et al., 2020; Khaki et al., 2020; Fan et al., 2022). Recent multimodal methods further improve prediction by jointly exploiting satellite imagery, vegetation indicators, and meteorological information (Lin et al., 2023; Zhang et al., 2025; Luo et al., 2026b).

Despite these advances, most existing approaches formulate yield prediction as a direct mapping from currently available observations to final yield. Such formulations treat vegetation signals primarily as observed predictors and do not explicitly model how crop states evolve after the observation cutoff. However, crop yield is the terminal outcome of a dynamic growth process, where future vegetation development contains additional information beyond the observed seasonal state. PhenoBIID differs from these approaches by explicitly modeling hidden vegetation evolution as an intermediate process between weather forcing and yield estimation.

2.2 Earth Observation World Models and Vegetation Dynamics

Earth observation modeling has shifted from static representation learning toward predicting environmental dynamics. World-model approaches aim to learn latent state transitions and forecast future Earth system states under changing conditions (Ha & Schmidhuber, 2018). In vegetation forecasting, existing studies explore weather-conditioned vegetation evolution, ecosystem simulation, and satellite-based trajectory prediction, demonstrating the potential of learning dynamic representations from multimodal Earth observation data (Benson et al., 2024; Iele et al., 2026; Luo et al., 2026a).

However, vegetation forecasting methods primarily evaluate future-state accuracy, while the downstream value of predicted states for agricultural decisions remains underexplored. Accurate trajectory reconstruction does not necessarily imply improved crop outcome prediction. PhenoBIID extends this perspective by connecting weather-conditioned vegetation state evolution with crop yield estimation, using predicted future vegetation trajectories as an explicit intermediate representation.

2.3 Agricultural Benchmarks and Evaluation Protocols

Large-scale agricultural benchmarks enable systematic evaluation of multimodal crop monitoring and yield prediction. Existing datasets combine yield statistics, satellite observations, and climate information, including SustainBench (Yeh et al., 2021), CropNet (Lin et al., 2024), CY-Bench (Kallenberg et al., 2026), CropClimateX (Höhl et al., 2026), and YieldSAT (Miranda et al., 2026). Long-term global yield products such as GDHY and GlobalCropYield5min further support large-scale agricultural analysis (Iizumi & Sakai, 2020; Cao et al., 2025).

Nevertheless, existing benchmarks mainly focus on final yield prediction accuracy and provide limited evaluation of whether models can forecast hidden vegetation development and whether such predictions improve downstream yield estimation. To address this gap, CropDynamicsBench provides an evaluation protocol that jointly assesses vegetation trajectory forecasting and yield prediction under different observation availability conditions.

3 Problem Formulation and Evaluation Protocol

We formulate in-season crop yield prediction as predictive world modeling. Given partial crop observations before harvest, we estimate final yield and test whether forecasting hidden vegetation evolution adds predictive value.

For each crop-grid-year sample, the seasonal growth process is represented as a sequence of crop-active temporal slots. Let V^{obs} denote the observed vegetation trajectory before the prediction cutoff, V^{future} the unobserved future trajectory, and W the corresponding meteorological conditions. The prediction problem can be expressed as:

$$(V^{\text{prev}}, V^{\text{obs}}, W, C) \to \hat{V}^{\text{future}} \to \hat{y},$$

where V^{prev} represents historical vegetation information, C denotes contextual variables, and $\hat{y}$ is the estimated annual yield. This formulation explicitly separates vegetation state forecasting from downstream yield prediction. In the main conditional evaluation, seasonal methods receive complete realized weather, including after the observation cutoff; this is not an operational issue-time availability claim.

3.1 Observation Cutoff Protocol

Crop calendars are used to align observations according to crop-active periods. Each sample is represented by at most twelve temporal slots corresponding to the growing season. To simulate different prediction stages, the final $\lceil rn \rceil$ active slots are treated as unobserved, where n is the number of active slots and r is the nominal unobserved fraction. We evaluate multiple cutoff settings ranging from fully observed trajectories to substantially incomplete growth processes.

3.2 Evaluation Objectives

The evaluation contains two complementary objectives.

Vegetation trajectory forecasting. Methods are evaluated on their ability to reconstruct the withheld vegetation states after the observation cutoff. This measures whether a model can capture future crop development under given environmental conditions.

Yield prediction. The completed vegetation trajectory is further used for annual yield estimation. This evaluates whether predicted future vegetation dynamics provide additional utility beyond directly using available observations.

This protocol enables a controlled analysis of the relationship between state forecasting and downstream agricultural prediction. Detailed dataset construction, preprocessing procedures, product descriptions, temporal partitions, and implementation specifications are provided in the appendices.

4 PhenoBIID: A Weather-conditioned Vegetation World Model

PhenoBIID connects weather-conditioned vegetation forecasting to annual yield through an explicit intermediate trajectory. Its latent state retains information about vegetation development, weather drives the transition, and observations provide feedback while available. After the observation cutoff, the model rolls forward using predicted feedback (Figure 2). The state representations are learned features, not explicit physiological crop variables.

Input representation and initialization. For each crop-grid-year sample, active months are ordered within the calendar year and packed into at most $K = 12$ slots; remaining positions are padded. For vegetation product p, an encoder initializes $S_{p,0} \in \mathbb{R}^{K \times d}$ from previous-year vegetation at the corresponding source months, validity, product support, and geographic/year context. The K latent tokens are retained, including positions with missing inputs. NDVI and GPP use separate state models. Data construction and exact encoding are specified in Appendices A and B.2.

Weather-conditioned state evolution. At active slot k, weather tokens W_k condition a transition operator $\mathcal{C}_p$. Vegetation is decoded as a residual to the standardized previous-year value $b_{p,k}$, set to zero when unavailable:

$$A_{p,k} = \mathcal{C}_p(S_{p,k-1}, W_k), \qquad \widehat{z}_{p,k} = b_{p,k} + h_p(\text{mean}\, A_{p,k}). \tag{1}$$

BIID uses bidirectional relevance modulation: each token set conditions weights on the other while retaining their separate token sets. Two synchronous interaction layers are followed by cross-attention and a gated interpolation with the incoming state. The detailed modulation equation, encoder, and gate are given in Appendix B.2.

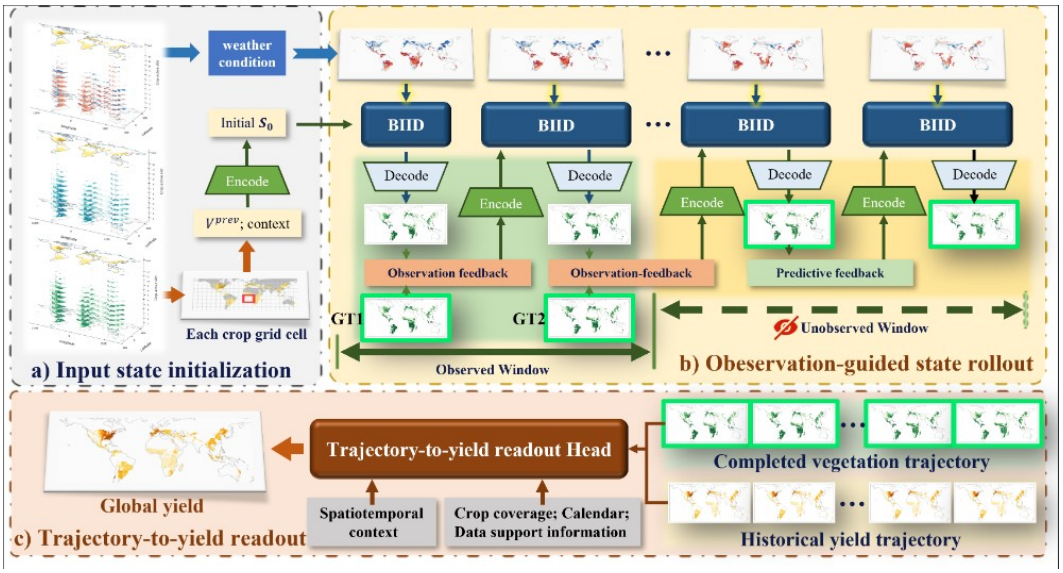

Figure 2: PhenoBIID uses weather-conditioned state transitions with observed feedback before the cutoff and predictive feedback afterward. The completed trajectory informs yield; readout details are shown in Appendix Figure 11.

Observation-guided rollout. Let $\mathcal{T}_r$ denote the withheld window. After predicting slot k, feedback uses a finite visible observation in standardized units, or the prediction otherwise:

$$u_{p,k} = \begin{cases} z_{p,k}, & k \notin \mathcal{T}_r,\ x_{p,k} \text{ finite}, \\ \widehat{z}_{p,k}, & \text{otherwise}, \end{cases} \qquad S_{p,k} = \mathrm{LN}(A_{p,k} + F_p(u_{p,k}))\,. \tag{2}$$

The projected scalar feedback is broadcast over state tokens; inactive slots preserve the state. Withheld vegetation never enters feedback. Predictions are converted back to physical units and replace only the unobserved window; missing values in the observed window remain missing.

Trajectory-to-yield readout. Completed-trajectory features are combined with history, weather, and cutoff-adjusted context C. For historical branches b with weights w_b, a crop-specific LightGBM correction g_b is added to a frozen historical anchor:

$$\widehat{y}^{(r)} = \sum_{b \in \mathcal{B}} w_b \left[\widehat{y}_b^{\,\mathrm{hist}} + g_b\Big(\widetilde{X}^{(r)}, C\Big) \right]. \tag{3}$$

When both products are used, their completed-trajectory features are concatenated. Each fitted readout is reused across observation cutoffs; the complete readout structure is shown in Appendix Figure 11, and feature encoding and branch definitions are provided in Appendices B.1 and B.4.

Staged learning. Vegetation dynamics and yield readouts are trained separately. State models minimize masked trajectory MSE with predicted feedback during training; visible-observation feedback is introduced at inference. Yield readouts are fitted on complete observed trajectories and receive completed trajectories at evaluation, with no end-to-end yield supervision of the state model. Historical predictions on readout-training rows are in-sample, so the terminal stack is not fully cross-fitted. Training boundaries and controls are documented in Appendices B.3 and D.5.

5 Experiments

5.1 Experimental setup

We evaluate annual yield and withheld vegetation on matched crop-grid-year samples from CropDynamicsBench. The main comparisons use thirteen evaluation years in expanding temporal blocks, with selection on preceding validation years and block-local refitting. Seasonal methods receive the observed vegetation window and complete realized weather; these are retrospective, conditional

Table 1: Mean annual yield RMSE (t/ha; mean ± sample SD across seeds) across unobserved windows. Lowest in bold; second-lowest underlined.

Unobserved	Method	Maize	Rice	Soybean	Wheat
10%	Random Forest	0.8509 ± 0.0011	$\underline{0.7699 \pm 0.0023}$	$\underline{0.3123 \pm 0.0003}$	0.8091 ± 0.0030
	LightGBM	$\underline{0.8459 \pm 0.0038}$	0.7826 ± 0.0033	0.3129 ± 0.0007	$\underline{0.7885 \pm 0.0053}$
	PHENOBIID	$\mathbf{0.8440 \pm 0.0031}$	$\mathbf{0.7545 \pm 0.0013}$	$\mathbf{0.2992 \pm 0.0037}$	$\mathbf{0.7639 \pm 0.0007}$
30%	Random Forest	0.8618 ± 0.0023	$\underline{0.7694 \pm 0.0003}$	$\underline{0.3134 \pm 0.0001}$	0.8060 ± 0.0023
	LightGBM	$\underline{0.8474 \pm 0.0036}$	0.7827 ± 0.0018	0.3147 ± 0.0002	$\underline{0.7892 \pm 0.0032}$
	PHENOBIID	$\mathbf{0.8464 \pm 0.0034}$	$\mathbf{0.7578 \pm 0.0010}$	$\mathbf{0.3024 \pm 0.0036}$	$\mathbf{0.7661 \pm 0.0012}$
50%	Random Forest	0.8713 ± 0.0012	$\underline{0.7749 \pm 0.0010}$	0.3223 ± 0.0003	0.8055 ± 0.0008
	LightGBM	$\underline{0.8552 \pm 0.0024}$	0.7908 ± 0.0032	$\underline{0.3207 \pm 0.0012}$	$\underline{0.7793 \pm 0.0035}$
	PHENOBIID	$\mathbf{0.8512 \pm 0.0034}$	$\mathbf{0.7610 \pm 0.0011}$	$\mathbf{0.3055 \pm 0.0029}$	$\mathbf{0.7717 \pm 0.0007}$
70%	Random Forest	0.8789 ± 0.0002	$\underline{0.7784 \pm 0.0009}$	$\underline{0.3234 \pm 0.0003}$	$\underline{0.8107 \pm 0.0017}$
	LightGBM	$\underline{0.8689 \pm 0.0006}$	0.7988 ± 0.0010	0.3243 ± 0.0008	0.8248 ± 0.0010
	PHENOBIID	$\mathbf{0.8584 \pm 0.0023}$	$\mathbf{0.7676 \pm 0.0021}$	$\mathbf{0.3118 \pm 0.0030}$	$\mathbf{0.7794 \pm 0.0010}$

comparisons, not operational issue-time forecasts. Seeds 42, 45, and 48 are used for the repeated main comparisons at 10%, 30%, 50%, and 70% unobserved windows.

Historical predictors use causal yield history and context; direct seasonal models additionally use observed vegetation and weather without forecasting vegetation. Completion controls hold the yield readout fixed and change only the hidden trajectory. Yield RMSE is computed spatially within each year and averaged with equal year weight; repeated results report seed mean and sample SD. Vegetation scores use withheld slots with valid references, separately by product. Full configurations and scoring are given in Appendices A.3 and B.6. The seasonal-weather substitution in Section 5.5 uses a separate nine-year evaluation described in Appendix C.6.

5.2 MAIN YIELD PREDICTION RESULTS

Table 1 compares PHENOBIID with Random Forest and LightGBM across observation cutoffs. The complete pipeline has the lowest seed-mean RMSE for each crop and listed window. Rice, soybean, and wheat improve in all three paired seeds; maize remains close to LightGBM and is initialization-sensitive. These are complete-method comparisons: by themselves, they do not isolate the contribution of forecasting vegetation. Table 11 reports the same pipeline at seed 42.

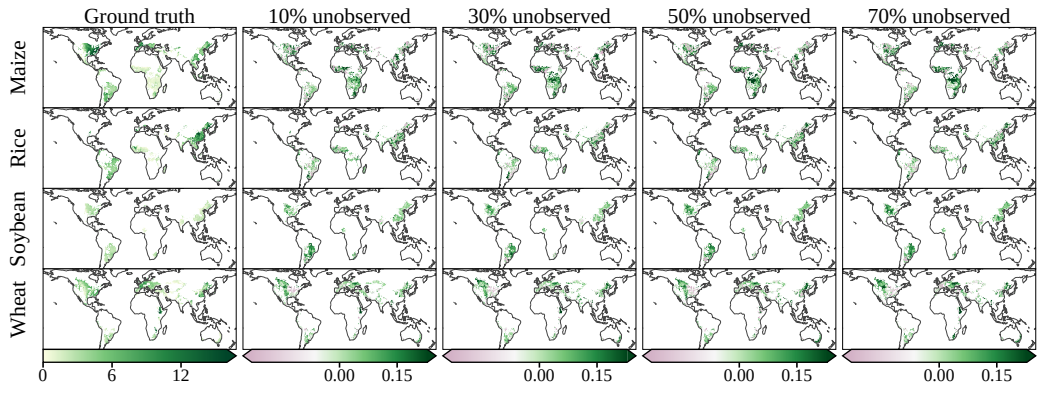

Figure 3: Ground truth and MAE reduction versus Random Forest or LightGBM from Table 1 (t/ha) at 10–70% unobserved windows. Positive values favor PHENOBIID.

Figure 3 shows the original paired spatial MAE reductions against the direct predictors. The full historical library, six-family direct comparison, and per-year results are retained in Appendix C.1. Seed and year-level consistency are reported in Appendices C.3.1 and C.4.

5.3 MECHANISM ANALYSIS

5.3.1 VALUE OF PREDICTIVE VEGETATION COMPLETION

We keep the observed window, weather, historical experts, and yield-readout weights identical, changing only the hidden-window completion rule. Figure 4 compares BIID with historical-mean vegetation completion. Both supply the same completed-trajectory interface. Predictive completion becomes more useful at longer unobserved windows for maize, soybean, and wheat, while rice favors historical-mean completion throughout. The added value is therefore crop- and cutoff-dependent, rather than a universal explanation for the complete pipeline's advantage.

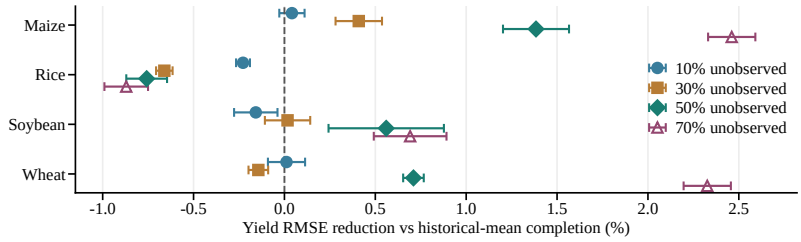

Figure 4: Mean RMSE reduction versus historical-mean vegetation completion under a shared readout. Error bars show seed SD. Positive values favor BIID.

Figure 5 illustrates the same-head mechanism at a 30% unobserved window. Forecasts below the historical vegetation mean lower the yield estimate in the maize, rice, and wheat examples; a forecast above the mean raises the soybean estimate. These four grid-years illustrate the mechanism, not typical improvement magnitudes. Previous-year and last-observation completion controls are retained in Appendix D.1.

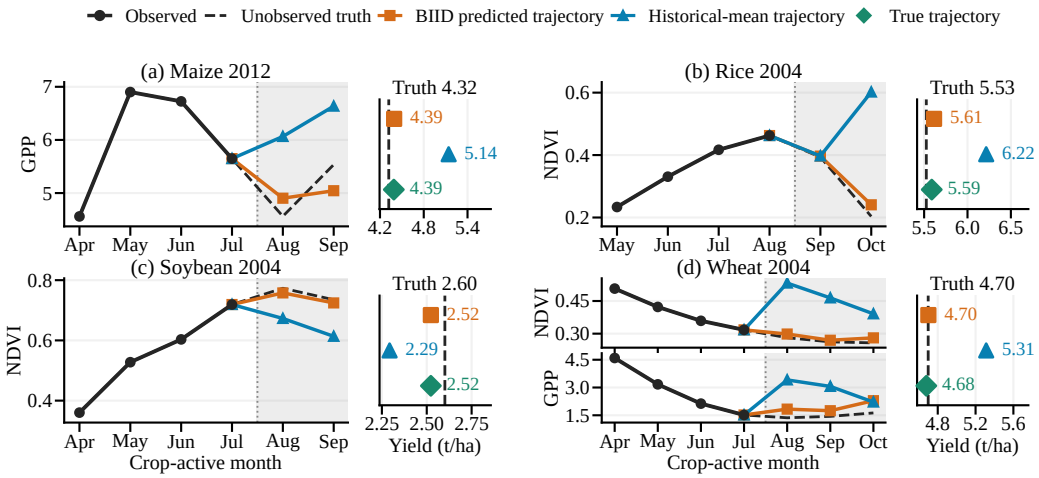

Figure 5: Case study at a 30% unobserved window, seed 42, shared readout. Gray marks the unobserved window. Left: vegetation trajectory. Right: yield from the BIID predicted trajectory, historical-mean trajectory, and true vegetation trajectory (t/ha); dashed line is truth.

5.3.2 VEGETATION FORECASTING ACCURACY AND YIELD UTILITY

BIID reduces vegetation RMSE relative to GRU in all fifteen reported crop–product–cutoff conditions. With the same frozen yield heads, BIID also has the lowest yield RMSE on all four crops at the 10% and 50% windows, and on maize, rice, and soybean at 30% (Tables 18–19). These controls change only the completion rule, retaining the seed-42 pipeline. Lower state error therefore does not guarantee greater yield value; anomaly correlations are reported in Table 20.

5.3.3 Role of weather-conditioned state evolution

This control removes weather variation from state evolution while retaining terminal weather and the same yield-head weights. The state model is retrained with all thirteen standardized weather channels fixed to zero, preserving the input dimensions and architecture (Appendix D.2). Figure 6 reports vegetation RMSE reductions of 2.58–19.94% from weather conditioning. Its downstream benefit is clearest at longer unobserved windows for maize, soybean, and wheat; rice does not show the same benefit. This isolates the state-weather pathway rather than removing weather from the whole yield system.

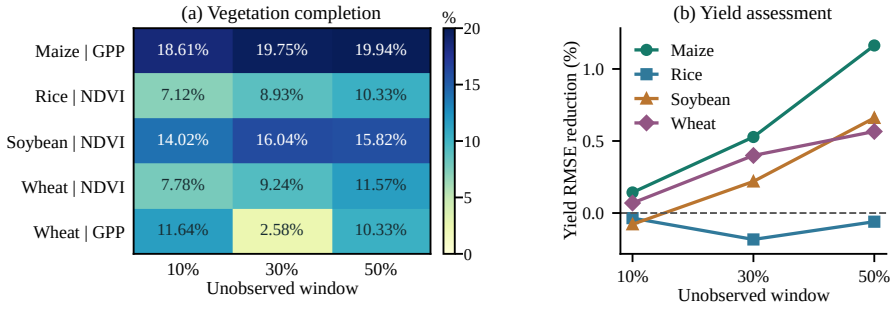

Figure 6: Mean annual reductions in (a) vegetation and (b) yield RMSE from weather-conditioned state evolution. Terminal weather and yield-head weights are fixed. Positive values favor weather conditioning.

5.3.4 Directional BIID ablation

We isolate the bidirectional interaction by replacing it with a parameter-matched pooled concatenation baseline or retaining only Weather→State or State→Weather modulation. Variants retrain only the interaction, retaining the same tokens, transition depth, frozen yield heads, and windows (Appendix D.6). Table 2(a)–(b) shows that Full BIID attains the lowest NDVI and GPP RMSE and the lowest frozen-head yield RMSE on all four crops.

Table 2: BIID interaction ablation, averaged over 10%, 30%, and 50% unobserved windows (seed 42). (a) Vegetation RMSE: NDVI or GPP (gC m^{-2} day^{-1}). (b) Yield RMSE (t/ha) with frozen readouts. Lower is better; best in bold, second-best underlined. Protocol details are in Appendix D.6.

	(a) Vegetation RMSE ↓		(b) Yield RMSE ↓			
Variant	NDVI	GPP	Maize	Rice	Soybean	Wheat
Concat	0.0503	0.5907	<u>0.8441</u>	<u>0.7569</u>	<u>0.2986</u>	0.7682
Weather→State	<u>0.0501</u>	<u>0.5859</u>	0.8445	0.7571	0.2988	0.7679
State→Weather	0.0509	0.5943	0.8451	0.7570	0.2987	<u>0.7676</u>
Full BIID	**0.0396**	**0.4611**	**0.8437**	**0.7567**	**0.2985**	**0.7674**

5.3.5 Value of vegetation inputs and product choice

Figure 7 tests observed vegetation inputs in direct LightGBM at a 10% unobserved window, using matched cohorts and training protocols (Appendix D.4). Retaining vegetation reduces yield RMSE across all four crops. Product choice also matters: NDVI+GPP performs best for maize and soybean, while NDVI alone performs best for rice and wheat. These controls establish the value of vegetation inputs beyond the proposed state-model architecture.

5.4 Robustness and independent regional evaluation

Paired spatial-block intervals favor PhenoBIID for soybean and wheat but span zero for maize and rice (Figure 15). Regional refits on independent yield labels outperform both direct predictors and historical-mean completion in five of six crop–country groups across all three cutoffs (Tables 16–17). Full analyses are in Appendix C.

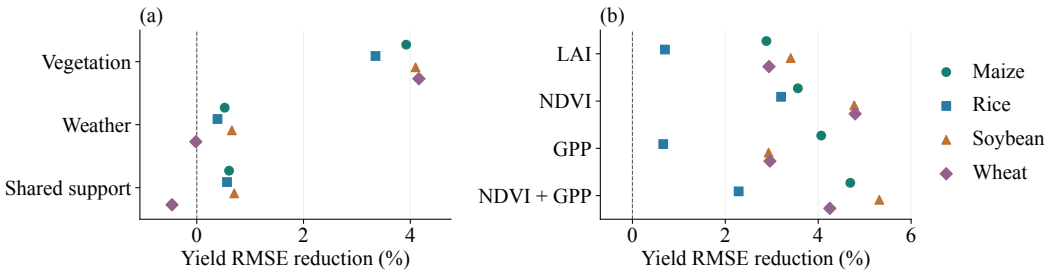

Figure 7: Direct LightGBM at 10% unobserved: (a) input groups versus matched removal; (b) products versus shared background. Positive values indicate lower yield RMSE.

5.5 Sensitivity to seasonal forecast weather

We substitute bias-corrected SEAS5 for unobserved-window weather over nine years (seed 42; 10%, 30%, 50%). Beyond-horizon slots use training climatology; direct baselines retain realized-weather training (Appendix C.6). Vegetation accuracy declines, but yield changes remain small (Table 3; Figure 8). This tests weather sensitivity, not prospective deployment.

Table 3: Relative yield RMSE (%) versus the named reference. vs ERA5 uses the same heads and realized weather; vs RF and vs LGB use Table 1 trees on the same years under realized weather. Positive values are worse. Cover: Appendix C.6.2.

	10% unobserved				30% unobserved				50% unobserved			
Crop	Cover	vs ERA5	vs RF	vs LGB	Cover	vs ERA5	vs RF	vs LGB	Cover	vs ERA5	vs RF	vs LGB
Maize	100%	+0.12	+0.74	+0.07	80%	+0.26	+0.31	+0.44	64%	+0.84	−0.30	+0.41
Rice	100%	−0.08	−1.42	−2.33	90%	−0.01	−0.80	−2.80	51%	+0.40	−1.03	−2.38
Soybean	100%	+0.06	−9.27	−8.73	94%	+0.59	−7.69	−7.81	83%	+0.89	−8.76	−8.05
Wheat	100%	+0.00	−6.85	−4.67	90%	+0.24	−5.76	−2.25	17%	+0.55	−4.62	−1.09

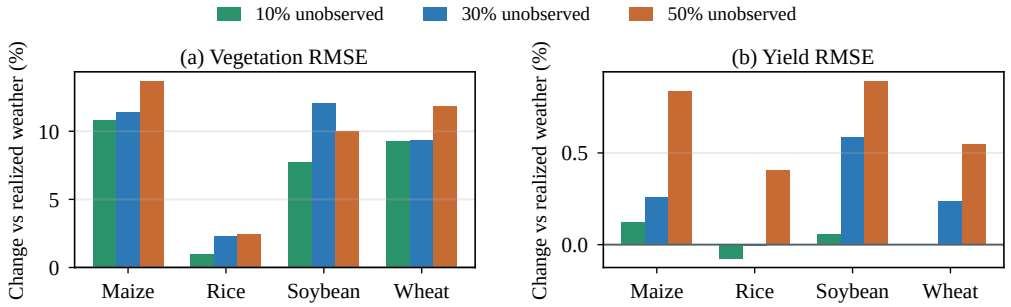

Figure 8: Relative RMSE change when unobserved-window weather is bias-corrected SEAS5 rather than realized ERA5-Land. (a) Vegetation (maize/wheat GPP; rice/soybean NDVI). (b) Yield. Positive values are worse than realized weather.

6 Conclusion

In this paper, we introduce PhenoBIID, a predictive crop world model that connects observation-guided, weather-driven vegetation evolution to yield through crop-specific readouts. To evaluate this framework, CropDynamicsBench aligns global four-crop records by crop-active periods for joint vegetation and yield assessment. Thirteen-year comparisons support its advantage over strong historical and direct predictors under given full-season weather. Input controls establish vegetation's value beyond the state architecture and support crop-specific product selection. Same-head controls show increasing completion benefits at longer unobserved windows for maize, soybean, and wheat. Ablations support weather-driven evolution and bidirectional interaction. Under SEAS5 substitution, yield accuracy remains close to the realized-weather reference in the evaluated setting. Together, these findings advance joint evaluation of growth forecasting and downstream yield value, providing a foundation for developing crop world models.

Reproducibility statement

Input schemas and training boundaries are specified in Appendix B. Repetition and regional protocols are provided with their results. Appendix B.7 records independent component reconstruction and complete evaluation, separately from raw-product preprocessing and public release.

AI use statement

Generative AI was used solely for language and grammar polishing and visualization assistance. The authors are responsible for independently verifying all code, claims, references, and artifacts. AI outputs were not treated as scientific evidence or cited as sources. The authors retain responsibility for verifying the cited literature and for the accuracy and integrity of the manuscript, including its scientific claims, numerical results, and references.

Ethics Statement

PhenoBIID and CropDynamicsBench investigate vegetation forecasting and crop yield prediction using Earth-observation products, meteorological data, and agricultural yield records, with potential applications in agricultural monitoring and food-security assessment. The reported experiments are retrospective, using realized weather as the primary input, and provide a benchmark for evaluating vegetation and yield prediction methods under observed conditions. Differences in data coverage, spatial aggregation, and crop-calendar alignment are key considerations for reliable evaluation across crops and regions. Model outputs can support agricultural decisions when used together with local validation, uncertainty assessment, and domain-expert oversight. Any redistribution of source-derived data must respect the corresponding providers' terms.

References

Weston Anderson, Shraddhanand Shukla, Jim Verdin, Andrew Hoell, Christina Justice, Brian Barker, Kimberly Slinski, Nathan Lenssen, Jiale Lou, Benjamin I. Cook, and Amy McNally. Preseason maize and wheat yield forecasts for early warning of crop failure. *Nature Communications*, 15: 7262, 2024. doi: 10.1038/s41467-024-51555-8.

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, June 2024.

Juan Cao, Zhao Zhang, Xiangzhong Luo, Yuchuan Luo, Jialu Xu, Jun Xie, Jichong Han, and Fulu Tao. Mapping global yields of four major crops at 5-minute resolution from 1982 to 2015 using multi-source data and machine learning. *Scientific Data*, 12(1):357, 2025. doi: 10.1038/s41597-025-04650-4.

Ziyue Chen, Yanzhao Wang, Xing Yan, Mei-po Kwan, Lin Yang, Qianqian Wang, Qiqi Zhu, Qiangyi Yu, Zhaozhong Feng, Bingbo Gao, Chaoqun Zhang, Yuheng Fu, Jianqiang Hu, Manchun Li, and Qiao Wang. An integrative framework for early estimation of global crop yields demonstrated under large-scale disruptions. *Nature Food*, 7:878–888, 2026. doi: 10.1038/s43016-026-01418-w.

Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P. Gomes. A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11873–11881, 2022. doi: 10.1609/aaai.v36i11.21444.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Adrian Höhl, Stella Ofori-Ampofo, Miguel-Ángel Fernández-Torres, Rıdvan Salih Kuzu, et al. A large-scale, multitask, multisensory dataset for climate-aware crop monitoring in the US from 2018–2022. *Scientific Data*, 13:72, 2026. doi: 10.1038/s41597-026-06611-x.

Andrew Hultgren, Tamma Carleton, Michael Delgado, Diana R. Gergel, Michael Greenstone, Trevor Houser, Solomon Hsiang, Amir Jina, Robert E. Kopp, Steven B. Malevich, Kelly E. McCusker, Terin Mayer, Ishan Nath, James Rising, Ashwin Rode, and Jiacan Yuan. Impacts of climate change on global agriculture accounting for adaptation. *Nature*, 642:644–652, 2025. doi: 10.1038/s41586-025-09085-w.

Irene Iele, Elena Mulero Ayllón, Paolo Soda, and Matteo Tortora. VegSim: A geospatial world model for scenario-conditioned vegetation simulation. *arXiv preprint arXiv:2606.21961*, 2026.

Toshichika Iizumi and Toru Sakai. The global dataset of historical yields for major crops 1981–2016. *Scientific Data*, 7(1):97, 2020. doi: 10.1038/s41597-020-0433-7.

David M. Johnson. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. *International Journal of Applied Earth Observation and Geoinformation*, 52:65–81, 2016. doi: 10.1016/j.jag.2016.05.010.

Michiel Kallenberg, Dilli Paudel, Stella Ofori-Ampofo, et al. CY-Bench: a comprehensive benchmark dataset for sub-national crop yield forecasting. *Earth System Science Data*, 18(6):3997–4018, 2026. doi: 10.5194/essd-18-3997-2026.

Endalkachew Abebe Kebede, Kevin Ong'are Oluoch, Stefan Siebert, Piyush Mehta, Sarah Hartman, Jonas Jägermeyr, Deepak Ray, Tariq Ali, Kate A. Brauman, Qinyu Deng, Wei Xie, and Kyle Frankel Davis. A global open-source dataset of monthly irrigated and rainfed cropped areas (MIRCA-OS) for the 21st century. *Scientific Data*, 12(1):208, 2025. doi: 10.1038/s41597-024-04313-w.

Saeed Khaki, Lizhi Wang, and Sotirios V. Archontoulis. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020. doi: 10.3389/fpls.2019.01750.

Muyi Li, Sen Cao, Zaichun Zhu, Zhe Wang, Ranga B. Myneni, and Shilong Piao. Spatiotemporally consistent global dataset of the GIMMS normalized difference vegetation index (PKU GIMMS NDVI) from 1982 to 2022. *Earth System Science Data*, 15(9):4181–4203, 2023. doi: 10.5194/essd-15-4181-2023.

Shunlin Liang, Jie Cheng, Kun Jia, Bo Jiang, Qiang Liu, Zhiqiang Xiao, Yunjun Yao, Wenping Yuan, Xiaotong Zhang, Xiang Zhao, and Ji Zhou. The global land surface satellite (GLASS) product suite. *Bulletin of the American Meteorological Society*, 102(2):E323–E337, 2021. doi: 10.1175/BAMS-D-18-0341.1.

Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, et al. 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*, pp. 5774–5784, October 2023. doi: 10.1109/ICCV51070.2023.00531.

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. Association for Computing Machinery, 2024. doi: 10.1145/3637528.3671536.

David B. Lobell, Adam Sibley, and J. Ivan Ortiz-Monasterio. Extreme heat effects on wheat senescence in India. *Nature Climate Change*, 2:186–189, 2012. doi: 10.1038/nclimate1356.

Chenxi Lu, Guoyong Leng, Lubin Han, Linfei Yu, Jiali Qiu, Lei Yao, Haiyang Tu, Xiaoyong Liao, Shengzhi Huang, and Jian Peng. Combining sun-induced chlorophyll fluorescence and seasonal climate forecast for 8-day dynamic in-season crop yield prediction. *Agricultural and Forest Meteorology*, 385:111221, 2026. doi: 10.1016/j.agrformet.2026.111221.

Junwei Luo, Shuai Yuan, Zhenya Yang, Yansheng Li, Zhe Liu, and Hengshuang Zhao. EO-WM: A physically informed world model for probabilistic earth observation forecasting. *arXiv preprint arXiv:2606.27277*, 2026a.

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 (CVPR)*, pp. 15870–15879, June 2026b.

Sara Minoli, Jonas Jägermeyr, Senthold Asseng, Anton Urfels, and Christoph Müller. Global crop yields can be lifted by timely adaptation of growing periods to climate change. *Nature Communications*, 13:7079, 2022. doi: 10.1038/s41467-022-34411-5.

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 (CVPR)*, pp. 22920–22930, June 2026.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, 284:107886, 2020. doi: 10.1016/j.agrformet.2019.107886.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B. 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.

Wenping Yuan, Shangrong Lin, and Tian Jie. A global gross primary production dataset based on rEC-LUE from 1982–2022, 2024. URL `https://zenodo.org/records/14350035`.

Jiaying Zhang, Kaiyu Guan, Zhangliang Chen, James Hipple, Yizhi Huang, Bin Peng, Sibo Wang, Xiangtao Xu, Zhenong Jin, Kejie Zhao, and Maxwell Jong. Aligning satellite-based phenology in a deep learning model for improved crop yield estimates over large regions. *Agricultural and Forest Meteorology*, 372:110675, 2025. doi: 10.1016/j.agrformet.2025.110675.

Appendix

Contents

A Dataset and Evaluation Details 14
A.1 Products, preprocessing, and cohort 14
A.2 Dataset Support and Evaluation Footprint 14
A.3 Shared evaluation design 17
A.4 Related datasets and prediction pathways 18

B Model and Baseline Implementation 18
B.1 Exact input and feature layout 19
B.2 State initialization, transition, and supervision 20
B.3 Temporal windows and checkpoints 21
B.4 History and terminal-head training 21
B.5 Evaluation, replay, and scope boundaries 23
B.6 Rolling Baseline Specifications 24
B.7 Implementation and Reproducibility 26

C Additional Experimental Results 27
C.1 Complete Thirteen-Year Results 27
C.2 Spatial Yield Assessment 31
C.3 Robustness and Completion Analysis 31
C.4 Paired Strong Direct-Model Repetitions 32
C.5 Independent Regional Statistical Labels 33
C.6 Seasonal Forecast Weather in the Unobserved Window 35

D Additional Analysis of PhenoBIID 36
D.1 State Architecture and Completion Controls 36
D.2 Weather in the State-Transition Pathway 38
D.3 Terminal Weather and Product-Support Paths 39
D.4 Input Value in Direct Seasonal Prediction 40
D.5 Historical-Anchor and Readout-Distribution Controls 41
D.6 Directional BIID Ablation 43

A Dataset and Evaluation Details

This appendix specifies the data sources, crop-grid sequences, sample support, and evaluation protocol used in Sections 3 and 5.1.

A.1 Products, preprocessing, and cohort

NDVI. The physical source is PKU GIMMS NDVI V1.2 consolidated data (Li et al., 2023), pinned to Zenodo record 8253971 and processed for 1982–2016. Native $1/12^\circ$ values are scaled by 0.001; codes above 1000 and fill code 65535 are excluded. Through 2002, accepted quality codes have method digit 1–5, AVHRR-quality digit 0, and MODIS digit 9; from 2003 the accepted code is 990. Each 6×6 native block is aggregated by a spherical-area-weighted valid-pixel mean and flipped to south-to-north latitude order. Monthly values are duration-weighted means of the two half-month means, using 15 days and the remaining calendar days, omitting nonfinite half-month means. This temporal mean is not additionally weighted by each half-month's valid-area fraction. Separate good-quality and physically available area–time fractions are retained. The consolidated product switches from AVHRR through 2002 to MODIS from 2003. Processed inputs are `ndvi_monthly`, `quality_fraction_monthly`, and `available_fraction_monthly` arrays under `Data/processed/pku_gimms_ndvi_v1p2/monthly_0p5`.

GPP. The source is monthly GLASS rEC-LUE GPP (Yuan et al., 2024), pinned to Zenodo record 14350035, also processed for 1982–2016. Native 0.05° unsigned-integer codes at most 60000 are valid and scaled by 0.01. A spherical-area-weighted 10×10 valid-pixel mean produces the 0.5° monthly total. Dividing by actual month length, including leap years, gives the scalar used here in $\mathrm{gC\,m^{-2}\,day^{-1}}$. The pipeline reads `gpp_daily_rate` and `valid_area_fraction` arrays under `Data/processed/reclue_monthly_gpp_v1/monthly_0p5`; it does not use an annual GPP total as a monthly trajectory. GPP is an ecosystem-model product driven by GLASS LAI, meteorology, and carbon dioxide. Its support fraction is provider-valid area divided by full grid area, not a retrieval-quality score or an independent measure of observation uncertainty.

Inherited support. GDHY annual yield (Iizumi & Sakai, 2020), ERA5-Land weather (Muñoz-Sabater et al., 2021), GLASS LAI support (Liang et al., 2021), and MIRCA-OS calendar/area metadata (Kebede et al., 2025) share the original sample identities. The original filter uses finite nonnegative yield, maximum monthly crop area above 100 ha, and at least three finite crop-active LAI months. World-cache target years start in 1982, while 1981 yield can supply history. Current-year LAI availability therefore affects inclusion even when the recipe contains only NDVI or GPP. No within-cell crop-purity mask is applied to these product values. The nearest MIRCA snapshot supplies both target-year calendar and crop-area context retrospectively; it is not an as-of-issue-time data vintage. Annual targets use the aggregate GDHY crop directories, rather than a new merger of major/second-season or spring/winter files. Sample counts and detailed calendar/cutoff rules are provided in Appendix A.2.

A.2 Dataset Support and Evaluation Footprint

A.2.1 Common samples and evaluation coverage

Table 4 reports the common evaluation population. All methods use the same crop–grid–year samples within each crop; adding a vegetation product does not expand this population. The yield, area, and LAI-availability eligibility criteria are specified in Appendix A.1.

Table 4: Grid–year samples in the thirteen-year evaluation. All methods use the same samples within each crop.

Crop	Grid cells	Evaluation samples
Maize	13,438	152,307
Rice	8,379	97,237
Soybean	5,599	68,317
Wheat	10,905	126,910

The physical record covers 1982–2016. Table 5 reports every inner-fitting, validation, full-fitting, and evaluation population. Earlier evaluation years can subsequently enter a later block's fitting window; the same rows are therefore not independent repetitions across fitting windows. Figure 9 shows the actual eligible footprint, rather than implying that every global grid cell or every retained cell has thirteen evaluation records.

Table 5: Grid–year counts by temporal block. Inner validation is merged after duration selection.

Crop	Fit end	Inner fit	Inner validation	Full fit	Evaluation
Maize	2001	185,538	20,632	206,170	36,489
Maize	2005	229,569	26,180	255,749	39,240
Maize	2009	281,923	26,133	308,056	76,578
Rice	2001	126,365	14,034	140,399	23,180
Rice	2005	155,502	16,161	171,663	24,216
Rice	2009	187,840	16,075	203,915	49,841
Soybean	2001	92,586	10,290	102,876	15,895
Soybean	2005	113,396	10,750	124,146	16,212
Soybean	2009	134,896	10,924	145,820	36,210
Wheat	2001	168,406	18,751	187,157	30,219
Wheat	2005	206,949	20,863	227,812	31,378
Wheat	2009	248,686	21,005	269,691	65,313

Table 6: Finite-value rates (%) among crop-active slots. Crop share is a growing-area proxy; disjoint denotes separated active months.

Crop	LAI	NDVI	GPP	Crop share	One interval	Disjoint
Maize	100.0	98.2	100.0	1.5	59.8	38.3
Rice	100.0	96.2	100.0	0.8	38.2	39.8
Soybean	100.0	97.9	100.0	2.2	67.5	32.4
Wheat	100.0	92.5	100.0	2.4	14.4	55.2

Availability in Table 6 means a finite value after aggregation within the inherited cohort. LAI monthly aggregation averages valid native pixels and temporal overlaps; a finite coarse value does not imply that every contributing pixel was valid or belonged to the target crop. Likewise, model-derived GPP can be spatially complete without providing an independent crop-specific observation. The growing-area/grid-area ratio is a retrospective maximum-monthly proxy, clipped to $[0, 1]$ in the input; it is not a crop-purity mask applied to the remote-sensing pixels.

A.2.2 Calendar alignment and observation cutoffs

MIRCA-OS snapshots for 2000, 2005, 2010, and 2015 supply crop-active months and crop-area context (Kebede et al., 2025). Each target year uses the nearest snapshot, even when that snapshot postdates the target year. Active natural months are sorted within the same calendar year and packed into at most twelve slots, with natural-month identities retained in the shared metadata. The sequence can contain gaps or multiple crop cycles; cross-year planting–harvest chronology is not reconstructed.

For a sample with n active slots, the last $\lceil rn \rceil$ slots are hidden at unobserved-window fraction $r \in \{0, 0.1, \ldots, 1\}$. Missing observations do not change this calendar-defined mask. At $r = 0$, all available current-year observations are retained; at $r = 1$, no current-year observation enters the observed window. The issue boundary is the end of the last retained active month, or the day before the first active month when the observed window is empty. Auxiliary lead-day metadata measures the distance to the final supplied active-month end, not to a verified harvest date. Product release latency is not modeled.

Each evaluation sample in Table 4 is one crop–grid–year. Across maize, rice, soybean and wheat, median active-slot counts are 6, 8, 5 and 11. At a 10% unobserved window, median withheld fractions of active slots are 16.7%, 16.7%, 20.0% and 16.7%, respectively. These are separate sample-wise medians: rounding a nominal ratio on a short trajectory makes the actual cutoff discrete, and a ratio of medians need not equal its median.

Calendar summaries describe the supplied support only. A contiguous interval excludes year-round activity; disjoint intervals contain a gap between active natural months. Neither class proves a single harvest or multiple harvests. In particular, activity on both sides of the calendar-year boundary can

reflect an unresolved cross-year season. Subset analyses based on these classes test sensitivity to the supplied calendar, not a ground-truth phenology annotation.

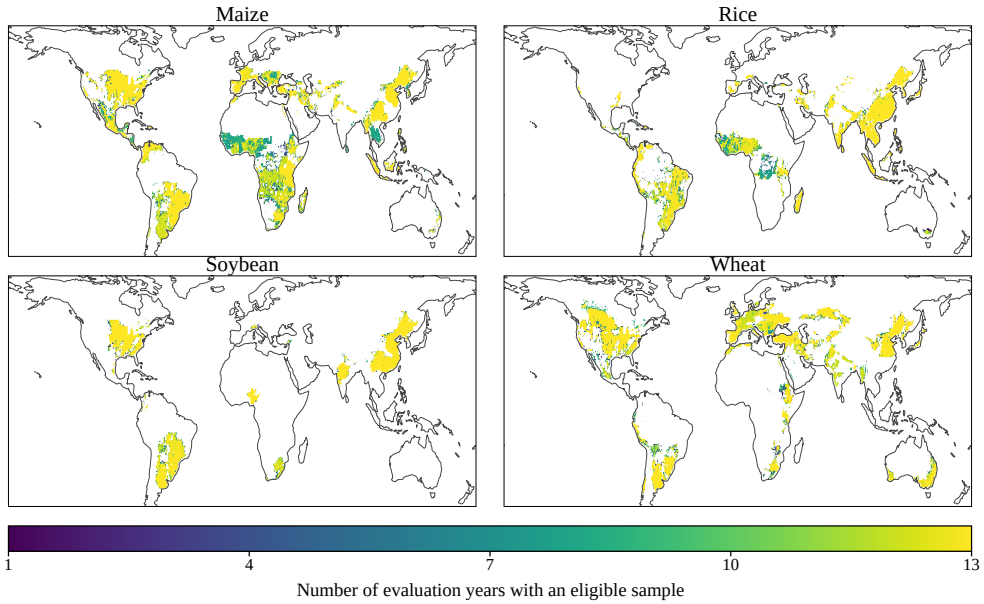

Figure 9: Eligible evaluation years per grid cell. Blank cells are excluded.

A.2.3 Area and calendar sensitivity

Table 7 changes the spatial weighting of the same predictions, not the fitting population or the historical comparator. With the growing-area proxy, the main pipeline's mean annual RMSE reductions are 9.50%, 3.16%, 16.59% and 8.49% for maize, rice, soybean and wheat. These are sensitivity estimates, not replacements for the grid-equal main score or estimates of error in total harvested production.

Table 7: RMSE reductions (%) versus the historical reference in Table 11 under alternative spatial weightings, 10% unobserved window.

Crop	Weighting	Main	Direct LGBM	Anchor control	Better years
Maize	Equal grid weights	+6.41	+6.04	+6.12	11/13
Maize	Crop-area weights	+9.50	+8.98	+8.21	12/13
Rice	Equal grid weights	+4.08	+0.78	+3.88	11/13
Rice	Crop-area weights	+3.16	+0.67	+2.48	7/13
Soybean	Equal grid weights	+9.24	+3.82	-5.86	12/13
Soybean	Crop-area weights	+16.59	+14.18	+2.11	13/13
Wheat	Equal grid weights	+5.21	+1.59	+4.03	10/13
Wheat	Crop-area weights	+8.49	+3.61	+6.64	11/13

Figure 10 partitions the inherited cohort using crop-share thresholds of 1% and 5%, or the calendar-support classes described above. Both predictions and the globally fixed historical reference are rescored within each subset. For wheat, mean annual RMSE reductions are 1.10%, 8.86% and 8.65% in the below-1%, 1–5% and above-5% subsets. Its three calendar classes also have positive average gains, showing that the pipeline retains predictive value across distinct crop-share and calendar-support groups. Soybean year-round support contains 46 grid–years over eight years; maize year-round support includes a year with only one grid cell. These sparse categories are retained with explicit sample counts so that the strength of support for each subgroup remains transparent.

A.2.4 Temporal coordinates with identical information

The supplied activity representation packs active months before padding, but already includes natural-month sine/cosine, calendar days and month gaps. To isolate the coordinate layout from information availability, we move each already-masked token to its source position in a twelve-month array. Its weather, vegetation, validity, support and calendar fields move together; inactive months remain

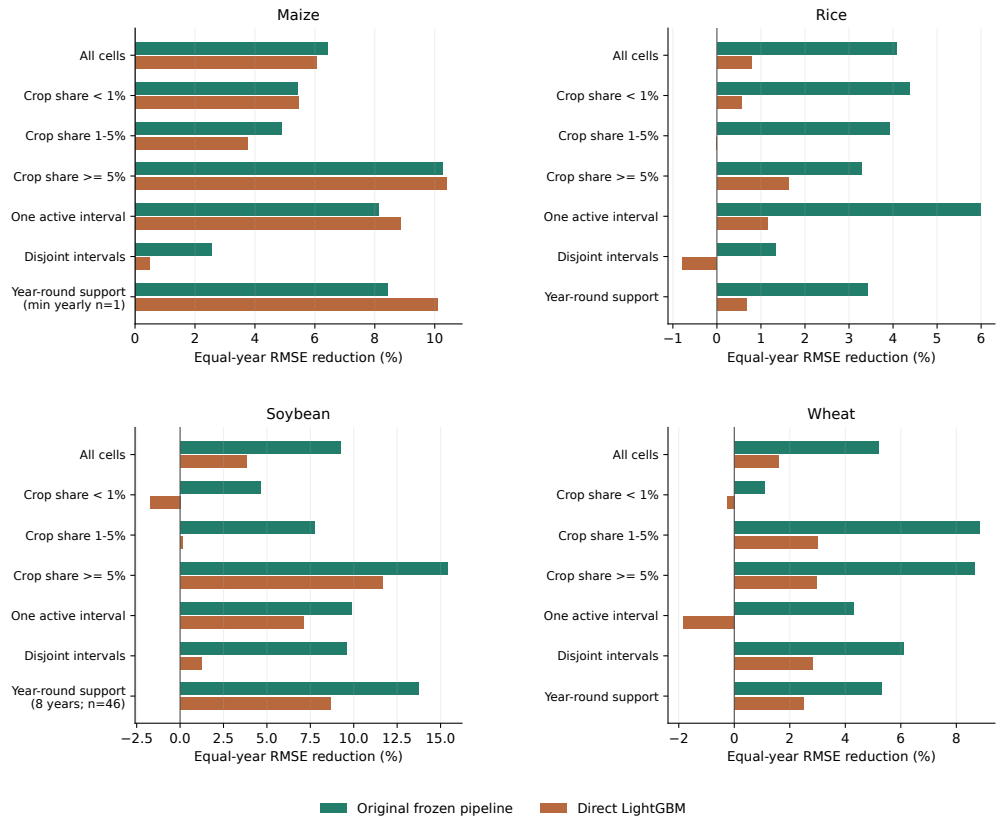

Figure 10: RMSE on fixed-support subsets at 10%, relative to the same historical configuration.

empty. Static summaries and the historical anchor are unchanged. The original active-slot cutoff is applied before rearrangement, and gathering the active tokens back reconstructs the original inputs exactly. Thus this comparison does not add non-active-month observations or repair cross-year harvest alignment.

We refit the direct CNN–RNN adaptation and Transformer at a 10% unobserved window, using the same architectures, training population, normalization, seed and preceding-validation epoch-selection budget. Both encoders explicitly mask intermediate padding. Twenty-four natural-month models are paired with twenty-four packed-coordinate models, matching normalization and parameter dimensions.

Table 8: Packed versus natural-month coordinates at a 10% unobserved window. Mean annual yield RMSE (t/ha). Lowest in bold; second-lowest underlined.

Model	Coordinates	Maize	Rice	Soybean	Wheat
CNN–RNN (adapted)	Packed active slots	<u>0.8704</u>	<u>0.8049</u>	<u>0.3377</u>	**0.8315**
CNN–RNN (adapted)	Natural-month coordinates	**0.8658**	**0.8005**	**0.3374**	<u>0.8327</u>
Transformer	Packed active slots	0.8993	0.8273	0.3521	0.8413
Transformer	Natural-month coordinates	0.8963	0.8289	0.3585	0.8420

Relative to natural-month coordinates, packing reduces CNN–RNN wheat error by 0.15%. Transformer packing improves rice, soybean and wheat, with its largest reduction of 1.78% for soybean. Because the information content is identical, these gains show that temporal coordinate layout can provide a useful architecture-dependent inductive bias.

A.3 Shared evaluation design

This section expands the setup in Section 5.1.

Temporal protocol. All global comparisons use the same crop–grid–year population and thirteen evaluation years (Tables 4 and 9). For each block, its preceding validation years select the crop signal, historical anchor, and component training settings, including epoch and tree counts. Historical MLP, TabM, state models, and terminal LightGBM are then freshly initialized and refitted through the block cutoff; weights remain fixed within that evaluation block. The corresponding evaluation outcomes do not enter selection. Earlier years may supply causal history or enter a later block's training and validation, but later choices never revise earlier predictions. Selection is performed separately by block, not by pooling validation windows; Appendix B specifies the internal and final fitting stages.

Table 9: Common temporal blocks for PHENOBIID and learned baselines. Validation years join the final fitting window after configuration selection.

Selection-training years	Validation years	Final-refit years	Evaluation years
1982–1999	2000–2001	1982–2001	2002–2004
1982–2003	2004–2005	1982–2005	2006–2008
1982–2007	2008–2009	1982–2009	2010–2016

Inputs, cutoffs, and repetition. Historical methods use past yields and fixed location/time features. Seasonal methods additionally receive the observed vegetation window and the same complete meteorological condition. The selected crop recipes are GPP for maize, NDVI for rice and soybean, and NDVI+GPP for wheat in all three blocks. An unobserved-window fraction r hides the last $\lceil rM \rceil$ of M active slots. We evaluate $r \in \{0, 0.1, \ldots, 1\}$. Zero is a fully observed reference, not early prediction; ratios are not universal days before harvest. The historical library and full sweep use seed 42. Three-seed comparisons repeat the selected configurations of the world model and two strong direct families at 10%, 30%, 50%, and 70% unobserved windows. Section 5.5 does not reuse this realized-weather, thirteen-year design; its substitution protocol is specified in Appendices C.6.1–C.6.3.

Metrics and reference models. Let $e^m_{c,t}$ be crop c's annual spatial yield RMSE for method m, in t/ha. We weight all thirteen years equally and report annual relative improvement against a named reference b:

$$\overline{E}^m_c = \frac{1}{13} \sum_{t \in \mathcal{T}} e^m_{c,t}, \qquad \overline{G}^m_c(b) = \frac{100}{13} \sum_{t \in \mathcal{T}} \left(1 - \frac{e^m_{c,t}}{e^b_{c,t}} \right). \tag{4}$$

We also report positive-year counts. Multi-seed tables give mean and sample SD of $\overline{E}$; their relative gains compare these seed-mean errors, not $\overline{G}$. For each crop, the lowest-$\overline{E}$ method among the 26 standard historical algorithms is the descriptive strong reference. This evaluation-ranked reference is used for descriptive reporting, not to select the seasonal predictor or its internal anchor. Its identity is fixed across years, cutoffs and seeds; all 26 algorithms are reported.

Baseline families. Table 11 groups the 26 historical algorithms into statistical, linear/robust, neighborhood/tree, and neural families. Learned methods use twenty causal history/context features; sequence models retain five ordered yield lags and validity. Project-specific two-stage residual variants and the HGB+MLP ensemble are outside this baseline library. Direct seasonal Ridge, Random Forest, LightGBM, GRU, Transformer, and an input-matched CNN–RNN adaptation instead predict yield without forecasting vegetation. They share masking, normalization rules, and a causal-trend anchor, with separate fits at each cutoff. Appendix B.6 gives configurations.

A.4 Related datasets and prediction pathways

Table 10 summarizes the dataset interfaces and prediction pathways discussed in Section 2.

B Model and Baseline Implementation

This appendix specifies the input schema, state transition, training boundaries, history-anchored yield readout, and baseline implementations. It documents the original saved model families rather than a newly trained end-to-end system.

Table 10: Related datasets (left) and prediction pathways (right).

A. Dataset interfaces

Dataset	Global yield grids	30+ yr vegetation record	Seasonal inputs
SustainBench	✗	✗	✓
CropNet	✗	✗	✓
CY-Bench	✗	✗	✓
California benchmark	✗	✗	✓
CropClimateX	✗	✗	✗
YieldSAT	✗	✗	✓
GlobalCropYield5min	✓	✗	✗
CROPDYNAMICSBENCH	✓	✓	✓

B. Prediction pathways

Method	Predict yield	Weather-driven vegetation	Predicted vegetation to yield	Multiple forecast times
MMST-ViT	✓	✗	✗	✗
Stage-aligned yield	✓	✗	✗	✗
PhenoYieldNet	✓	✗	✗	✓
NDVI-peak fitting	✓	✗	✓	✓
Contextformer	✗	✓	✗	✗
VegSim (preprint)	✗	✓	✗	✗
EO-WM (preprint)	✗	✓	✗	✗
PHENOBIID	✓	✓	✓	✓

✓: provided or evaluated; ✗: not part of the cited interface. Seasonal inputs organize imagery or vegetation through growing-season windows or crop calendars. The 30-year criterion concerns paired vegetation records, not yield history alone. Predicted vegetation includes reconstruction and extrapolated summaries. Multiple forecast times indicate yield prediction evaluated at different points within the growing season.

B.1 EXACT INPUT AND FEATURE LAYOUT

Observed-window completion. For product p, let $x_{p,k}$ be its physical reference value and $\mathcal{T}_r$ the unobserved active-slot window. The terminal encoder receives

$$\widetilde{x}_{p,k}^{(r)} = \begin{cases} x_{p,k}, & k \notin \mathcal{T}_r, \\ \widehat{x}_{p,k}^{(r)}, & k \in \mathcal{T}_r, \end{cases} \qquad k \text{ active.} \tag{5}$$

Missing observations in the observed window remain missing in this completed trajectory; finite predictions make unobserved slots available to the feature encoder. Hidden reference values are used for evaluation, not feedback.

For a batch of B samples, weather is $W \in \mathbb{R}^{B\times K\times V}$, previous values are $X_p^- \in \mathbb{R}^{B\times K}$, and the latent state is $S_{p,k} \in \mathbb{R}^{B\times K\times d}$, with $K = 12$, $V = 13$, and $d = 128$. Each scalar state branch receives weather of shape $B \times 12 \times 13$, previous values, previous validity, previous support, and active-slot validity each of shape $B \times 12$, and context of shape $B \times 5$. These are the six registered forward inputs. Yield history, current-year remote quality, crop-area fraction, and yield targets are not state inputs. Natural-month indices select source arrays but are not an additional input to this state network, which uses learned packed-slot positions. The weather-variable order is `d2m`, `t2m`, `stl1`, `stl2`, `swvl1`, `swvl2`, `swvl3`, `ssrd`, `pev`, `u10`, `v10`, `sp`, and `tp`.

The terminal design matrix has the following zero-based, half-open slices:

Slice	Width	Content
$[0, 20)$	20	15 yield-history features plus 5 context features
$[20, 309)$	289	Crop-area fraction plus 12×24 metadata fields
$[309, 465)$	156	12×13 standardized actual weather values
$[465, 501)$	36	NDVI or GPP trajectory encoding
$[501, 537)$	36	GPP encoding in the NDVI+GPP recipe only

Thus single-product heads have 501 inputs and the concatenated head has 537. The common block is 465-dimensional; historical expert predictions are additive baselines, not extra columns in that block.

Metadata is not history. Within each 24-field slot, offsets 0–4 are activity, natural-month sine and cosine, calendar days divided by 31, and the gap from the previous packed month divided by 12. Offsets 5–10 are LAI validity, NDVI validity, NDVI good area–time fraction, NDVI available area–time fraction, GPP validity, and GPP valid-area fraction. Offsets 11–23 are the thirteen weather-validity flags. All recipes retain this shared metadata, including other products' observed-window support. Only offsets 5–10 in unobserved window slots are replaced: each is imputed by its training grid/month average, falling back to the training global month average, then the overall active-slot average if necessary. History/context, crop area, calendar fields, weather validity, and weather values remain unchanged. This is not a product-pure modality ablation.

Normalization and trajectory encoding. Terminal NDVI mean/standard deviation come from finite active training slots in the stable observed cache; GPP statistics are fitted analogously in the signal cache. State normalization is fitted separately on its own training window and saved with its checkpoint. State outputs are first converted back to physical units, then restandardized using the terminal encoder's saved product scale; the two normalization objects are not interchangeable. Weather is standardized variable-wise on active finite training values; missing normalized inputs are zero-filled. Product arrays retain physical missingness until encoding.

Each product contributes twelve standardized values and six summaries (mean, population standard deviation, maximum, minimum, slot sum, and zero-based peak-slot index divided by 11), followed by the same eighteen features for local anomalies. Anomalies subtract a training grid/natural-month mean, falling back to the training global mean for that month. An absent fallback becomes zero in standardized units. Training anomaly construction excludes the row's own grid/month observation, but can use later years within the fitting window; it is leave-one-record-out, not forward-in-time cross-fitting. Evaluation uses the full fitting-window historical mean. Missing and inactive entries contribute zero to encoded slot values and are excluded from summaries. In the available observed-window interface, a finite generated unobserved window is valid regardless of the withheld reference's missingness; missing observed-window values are not filled in the terminal trajectory. The slot sum is a sum of standardized values, not a day-weighted seasonal integral of GPP.

B.2 State initialization, transition, and supervision

Previous-year initialization. Both products are gathered at the sample's grid cell and target-year source months from calendar year $t-1$, not from a separately reconstructed preceding crop cycle. Previous values are normalized with current branch training statistics. Their masks require finite values and active slots; missing values and support are zeroed. Previous NDVI support is its quality fraction, whereas previous GPP support is its valid-area fraction. Since both products start in 1982, the 1982 target has no previous product observations and is not backfilled from later years.

At each of twelve positions, value, validity, and support are linearly projected from three channels to width 128 and augmented with a learned position. An always-valid null token is appended. A one-layer pre-normalized Transformer encoder uses four attention heads, feed-forward width 256, GELU, and dropout 0.1, masking missing observation keys. The first twelve outputs receive a projected five-dimensional context, layer normalization, and a learned product identity. All twelve resulting latent tokens are retained, including outputs at missing/padded input positions; the null token is not retained. Consequently mean readout is over twelve latent tokens, not just the available observations.

Bidirectional relevance modulation. The reference transition uses bidirectional relevance modulation (BIID) to condition vegetation and weather representations on one another while retaining their separate token sets. For source tokens S and target tokens T, one direction computes

$$\alpha_j(S,T) = \mathrm{softmax}_j\left(\frac{1}{n_S\sqrt{d}}\sum_{i=1}^{n_S} Q(S_i)^\top K(T_j)\right), \qquad \mathcal{M}(S,T)_j = O(\alpha_j(S,T)V(T_j)). \quad (6)$$

Here n_S is the source-token count and Q, K, V, O are learned linear maps. Source-dependent weights modulate target-side values. Both directions update synchronously, followed by residual feed-forward blocks. Two such layers are followed by cross-attention and a learned interpolation with the incoming state. Appendix D.1 compares it with a GRU and other completion rules.

Recursive update. Weather values are individually projected from one channel to width 128 and augmented by learned variable identity, packed-slot position, and context, then layer-normalized. Each active step supplies thirteen weather tokens to two synchronous BIID layers with two-layer width-256 GELU feed-forward blocks and dropout 0.1. Following relevance modulation, four-head attention queries the modified weather with the modified state. Its output is added to the modified state to form a candidate. A sigmoid linear projection of incoming state concatenated with candidate interpolates between them, followed by layer normalization. The scalar readout is layer normalization, linear $128 \to 64$, GELU, and linear $64 \to 1$, added to the valid previous-year value for that slot. A linear $1 \to 128$ feedback message is broadcast over all twelve state tokens and layer-normalized. Missing observations in the observed window use predicted feedback; finite observed-window values

replace that feedback only after the current prediction. Inactive steps preserve the state. No clipping of physical NDVI or GPP predictions is applied.

Loss and selection. Training minimizes squared error in standardized product units, summed over finite active targets and divided by their count within each minibatch. This loss does not weight by crop area, monthly growing-area weights, or product support. Current observations supervise the loss but are never fed back during training: recursion uses the model's own predictions. There is no observed-window-specific training or fine-tuning. AdamW uses learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 256, gradient norm clipping at 1, and CUDA bfloat16 autocasting. Inner selection allows at most 30 epochs with patience 5 and minimizes the mean of the two inner years' physical state RMSEs. A new model is then initialized with the same seed, statistics are refitted on the full fitting window, and training runs for exactly the selected number of epochs. The seasonal runner loads this full-refit `model.pt`, not the inner-selection checkpoint.

B.3 Temporal windows and checkpoints

Let o be the development origin and $\tau = o - 3$ the fitting cutoff. State inner fitting ends at $\tau - 2$; inner selection uses $\tau - 1, \tau$; full state refitting and terminal-head fitting use target years 1982 through τ. The windows are:

Origin	Inner fit end	Inner selection	Full fit end	Evaluation years
2004	1999	2000–2001	2001	2002–2004
2008	2003	2004–2005	2005	2006–2008
2012	2007	2008–2009	2009	2010–2016

The last origin supplies the seven-year final block. Models, normalization, historical means, and support-imputation statistics remain fixed at their registered windows; previous-year products and yield-history inputs roll forward annually. The three blocks jointly form the main thirteen-year evaluation. For every trainable component, configuration selection uses only the two inner-selection years shown above, followed by fresh initialization and full-window refitting through τ.

The broad main comparison uses seed 42; multi-seed results additionally use seeds 45 and 48 under the same selected crop recipes. The saved epoch counts below are ordered by origins 2004, 2008, and 2012. They were selected on the preceding inner years and specify the duration of the subsequent full refit; they are not performance measurements.

Crop	NDVI state	GPP state	History MLP	History TabM
Maize	6, 14, 14	18, 5, 15	27, 3, 2	Not used
Rice	4, 7, 28	22, 7, 20	10, 8, 13	Not used
Soybean	9, 11, 15	12, 6, 13	5, 2, 1	0, 37, 0
Wheat	9, 8, 11	14, 5, 14	3, 1, 3	Not used

The selected soybean TabM duration is zero at origins 2004 and 2012. Those freshly initialized corrections receive no gradient update; the underlying historical MLP is still fully refitted.

Internal selection and final refitting. Internal-fitting cutoffs are 1999, 2003, and 2007. When selecting TabM duration or terminal LightGBM tree count, validation-row history predictions use the corresponding internal-window models. Feature scaling, quantile bins, historical means, support imputation, and residual statistics also use only internal-fitting rows. After selection, all components are freshly initialized and refitted in dependency order through 2001, 2005, or 2009, with statistics recomputed on the full fitting window. Final-refit models and statistics therefore do not construct the internal-validation inputs. The selected recipe and training durations are shared by the seed repetitions. The thirteen-year score combines the three blocks' predictions.

B.4 History and terminal-head training

Historical comparison inventory. The main thirteen-year comparison includes 26 standard historical algorithms. The historical comparator is distinct from the additive history anchor inside the

seasonal yield head. Project-specific two-stage residual configurations and the HGB+MLP ensemble are excluded from this standard-algorithm library. The historical components retained inside the proposed model are specified separately in Table 13. Model definitions and refitting windows appear in Appendix B.6.

Historical features. The fifteen features are five normalized calendar-year yield lags and five availability flags, historical mean, extrapolated historical linear trend, historical standard deviation, historical trend slope, and observation-count fraction. Mean and trend are centered/scaled with training target statistics; standard deviation and slope are divided by the target scale. Missing lags are zero with mask zero. Grid-specific moments are updated only after that year's features have been emitted. The trend uses at least five observations, otherwise the historical mean, and is clipped below at zero. The historical window can include 1981. In this terminal cache the count denominator is the full source-year span, 35; it is a fixed time coordinate, not a future yield statistic. The five context fields are sine/cosine of latitude, sine/cosine of longitude, and $(t - 1981)/35$. Later evaluated years use preceding evaluated years' observed yields as lags and in historical moments. This assumes annual availability of those retrospective yields; release delays are not modeled.

Learned historical MLP. The historical MLP takes these twenty history/context inputs and has widths $20 \to 256 \to 128 \to 1$, GELU activations, dropout 0.1, and a zero-initialized final layer. Its standardized residual is added to the causal trend baseline using training-fitted residual mean and scale. For each temporal block, an inner model is fitted through $\tau - 2$ with yield-residual MSE only, AdamW at 3×10^{-4} with weight decay 10^{-4}, gradient clipping at 1, batch size 2048 accumulated in microbatches of 256, and bfloat16 autocasting. At most 50 epochs are considered, with patience 8. Mean annual physical-yield RMSE on years $\tau - 1, \tau$ selects the epoch count. A fresh MLP is then initialized and refitted on all target years through τ for exactly that duration.

Soybean historical expert. Soybean adds a numerical-embedding TabM residual to its saved historical MLP. Its twenty history/context inputs use a training-fitted StandardScaler and training-only quantile bins (up to sixteen), with 12-dimensional piecewise linear embeddings. TabM has sixteen ensemble members, three width-128 blocks, and dropout 0.1; member outputs are averaged for prediction. It fits standardized residuals to the MLP with member-wise MSE, AdamW at 10^{-3}, weight decay 10^{-4}, batch size 1024, microbatch size 256, FP32, clipping at 5, at most 60 epochs, and patience 8. Mean annual yield RMSE on the same preceding validation years selects the duration, including epoch zero. A fresh TabM component is then refitted through τ for the selected duration. The seasonal expert uses the resulting uncalibrated `component_prediction`, not the source experiment's subsequent calibrated prediction.

Readout notation. Equation 3 suppresses crop and signal-configuration indices and absorbs residual scaling into the correction function. With Φ denoting the feature encoder specified in Appendix B.1,

$$g_b(\widetilde{X}^{(r)}, C) = \mu_b + \sigma_b f_b\Big(\Phi(\widetilde{X}^{(r)}, C)\Big),$$

where f_b is the fitted LightGBM predictor in standardized residual units. Thus g_b outputs a correction in t/ha; it does not operate directly on unencoded trajectories. Each crop and signal configuration retains its own fitted readout and the original historical branches.

Crop-specific residual heads. Maize averages two corrected predictions with fixed weights $(0.5, 0.5)$: one anchored to the causal trend and one to the historical MLP. Rice and wheat use one MLP-anchored correction; soybean uses one MLP-plus-TabM-anchored correction. Every correction is a LightGBM squared-error regressor of the complete feature matrix. Each branch's residual mean and standard deviation are fitted on training rows (standard deviation floored at 10^{-6}) and retained across product recipes. The registered trees use learning rate 0.03, up to 1200 trees, 15 leaves, minimum child samples 200, L2 penalty 10, row and column subsampling 0.9, and row-subsampling frequency 1. Training weights are $N/(JN_y)$ for a row in year y, where N is total rows, J the number of training years, and N_y that year's rows. Each training year therefore has equal total weight. Mean annual RMSE on years $\tau - 1, \tau$ selects the tree count with patience 60. A fresh LightGBM readout is then refitted on all rows through τ using that count. Selected tree counts vary by crop, origin, branch, and recipe and are recorded in each head configuration.

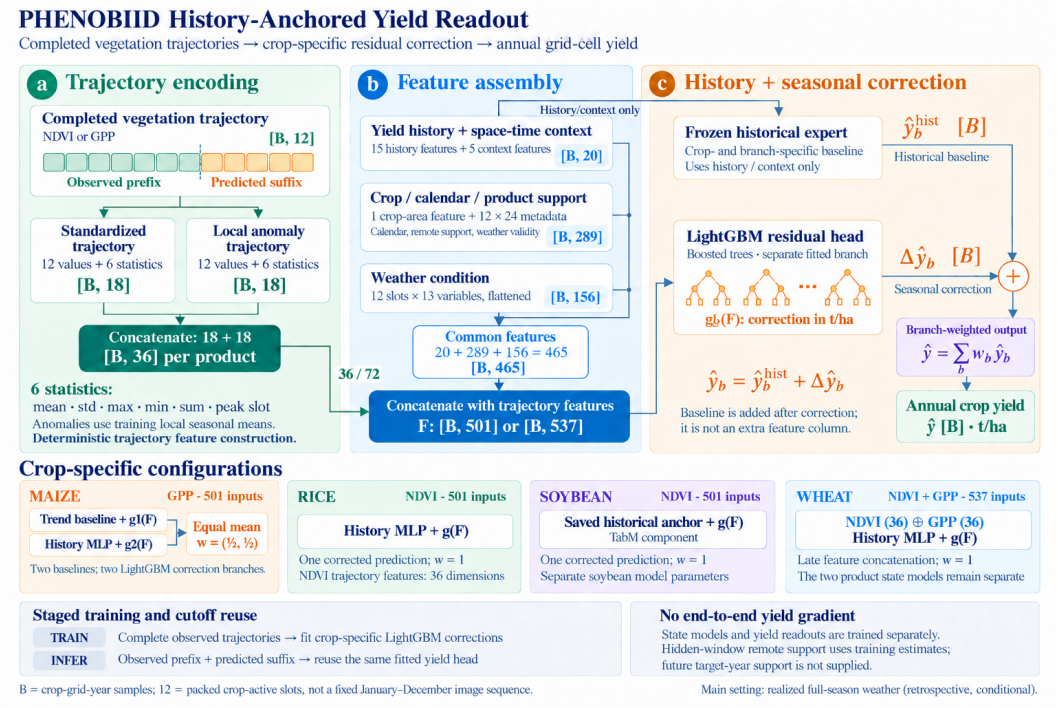

Figure 11: Detailed trajectory-to-yield readout. Each completed NDVI or GPP trajectory is encoded by standardized and local-anomaly summaries, concatenated with history, space–time context, crop/calendar/data-support metadata, and weather, and passed to a crop-specific LightGBM residual head. The resulting correction is added to a frozen historical anchor; crop-specific branches are combined with fixed weights. State models and yield readouts are trained in separate stages.

NDVI configurations are registered in the observed-head inventory; GPP and NDVI+GPP configurations are registered in `crop_head_signal_match_v1` and `crop_head_signal_pair_v1`. Within every temporal block, distinct trees are freshly refitted on their respective complete observed-product features under the preceding-validation protocol above. Each resulting head is reused across the seasonal cutoff sweep.

History-training caveat. Raw lag features are past-only, but learned history predictions for terminal training rows are in-sample replays of the corresponding fitting-window historical checkpoint. Internal selection uses the internal-fitting window; final refitting uses the full fitting window. This in-sample training-row reuse does not extend to the internal-validation rows, whose upstream models exclude those years as specified in Appendix B.3. They are not out-of-fold or strictly forward predictions. For soybean, this also applies to the added TabM correction. Consequently residual-head targets can reflect optimism in the fitted history expert. Complete observed trajectories and their support train the heads, while generated trajectories and imputed support are substituted only at evaluation. This distribution change precludes describing the terminal stack as fully out-of-fold and forecast-matched.

B.5 Evaluation, Replay, and Scope Boundaries

Historical reference in archived sweeps. The preceding NDVI reuse experiment attached a `strong` prediction selected from fifteen saved historical configurations, allowing epoch-zero candidates and applying no prediction calibration. The present main table does not use that cached field as its comparison denominator: it recomputes the 26 standard algorithms over the thirteen-year population. A single historical algorithm is then held fixed across years and cutoffs within each crop. This change affects the reported reference, not the frozen seasonal predictions or their blockwise refitted historical anchor.

Seasonal sweep. The top-level runner replays cached NDVI observed-window trajectories and computes GPP trajectories with its separate frozen state checkpoint. It evaluates both state and

historical-mean completion for each recipe, plus NDVI-state/GPP-historical-mean completion for the paired recipe. A common unobserved-window mask is formed from $\lceil rn - 10^{-9} \rceil$ hidden active slots, clipped to $[0, n]$, for $r \in \{0, 0.1, \ldots, 1\}$. Missing observations do not change this calendar-defined cutoff. At $r = 0$, no forecast value enters the completed trajectory; at $r = 1$, no current-year observation enters feedback. The code checks sample/target identities, input and checkpoint hashes, exact replay of complete-observation development predictions for all three recipes, and exact replay of the cached NDVI seasonal yield predictions. These checks establish consistency with saved artifacts, not causal data availability or a fresh native-source reconstruction.

Yield RMSE is unweighted across grid cells within each crop/year. Summaries average annual quantities with equal year weight. Mean annual percentage improvement, the mean of $100(1 - \mathrm{RMSE}_y/\mathrm{RMSE}_y^{\mathrm{ref}})$, is distinct from the percentage improvement of mean annual RMSE; the summary code records both. State RMSE on the unobserved window uses only hidden slots whose physical reference is finite. Median lead days are derived from signed natural-month indices and calendar boundaries, not from observed harvest dates or a constant conversion from unobserved-window percentage to days.

What extends beyond issue time. Both the state rollout and direct weather columns condition on actual post-issue weather. The cohort uses full-year LAI availability, and the calendar and crop-area metadata can come from a later MIRCA snapshot. Retrospective NDVI and model-derived GPP have no enforced release-date or data-vintage cutoff. The loader reads complete reference arrays for replay and scoring; the narrower boundary is that withheld remote values do not enter observed-window feedback or completed unobserved-window features, and their six unobserved-window support fields are replaced by training estimates. Observed-window support from all products remains visible. Earlier evaluated yields and previous-year product fields are supplied annually, without checking their publication latency. These conditions define the additional checks required for operational issue-time forecasting. Accordingly, the reported experiments are framed as retrospective conditional prediction. NDVI and GPP use separate product-state weights, while GPP remains model-derived and shares vegetation and climate dependencies. This scope positions the benchmark as an evaluation of weather-conditioned state completion and downstream yield value, rather than a causal intervention or season-resolved harvest simulator.

B.6 Rolling Baseline Specifications

B.6.1 Historical library

Each crop and temporal block is fitted independently using Table 9. All learned historical predictors receive the same fifteen lag/summary features and five location/time features described in Appendix B.4. Feature normalization and residual center/scale are recomputed from the corresponding fitting rows. The target is a standardized residual to the causal per-grid trend, which is added back before scoring. Targets from the evaluation years within a block do not enter that block's new-baseline fitting or checkpoint selection. Earlier evaluation years can become fitting data for a later block. No negative predictions are clipped in this comparison.

The seven statistical methods are global training mean, causal grid historical mean, previous-year yield, three- and five-year moving means, exponential smoothing, and per-grid linear trend. Exponential smoothing selects $\alpha \in \{0.1, 0.2, 0.35, 0.5, 0.65, 0.8, 1\}$ on the preceding internal validation years. Missing histories use training-period fallbacks. Historical moments are constructed from the original yield population before indexing the shared vegetation cohort.

The six linear/robust models are ordinary least squares, Ridge ($\alpha = 10$), Lasso ($\alpha = 10^{-3}$), Elastic Net ($\alpha = 10^{-3}$, L1 ratio 0.5), Huber regression ($\epsilon = 1.35$, $\alpha = 10^{-4}$), and linear SVR ($C = 1$). They use training-fitted feature standardization and double precision. Approximate KNN uses 25 inverse-distance-weighted neighbors in a training-standardized HNSW index, with construction/search parameters 100/96. The decision tree has maximum depth 20 and minimum leaf size 16. Random forest and Extra Trees each use 128 trees, maximum depth 20, minimum leaf size 8, and feature fraction 0.75.

Histogram gradient boosting (HGB) uses 200 iterations, learning rate 0.05, 31 leaves, and L2 penalty 1. LightGBM uses learning rate 0.03, 31 leaves, minimum leaf size 40, row/feature fractions 0.9, and L2 penalty 1. XGBoost uses learning rate 0.03, depth 6, child weight 8, row/feature fractions 0.9, and

L2 penalty 1. The latter two select the tree count from at most 1,200 iterations with 60-round early stopping on pooled internal validation RMSE, then refit all rows through the block cutoff. Classical methods otherwise retain their fixed original-suite settings.

The six neural models are MLP, RNN, LSTM, GRU, TCN, and Transformer. Five yield lags and their indicators form a 5×2 sequence; ten summary/context features enter the output head. Recurrent encoders have one 64-wide layer. The TCN has two 64-channel convolutions of kernel width 3, with dilations 1 and 2. The Transformer has two 64-wide layers, four attention heads, and feed-forward width 128. The MLP flattens the lag sequence. All use a 128/64-wide output MLP with dropout 0.2, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 4,096, gradient clipping at 1, at most 40 epochs, and patience 6. Uniform-row residual MSE is optimized; mean annual internal validation RMSE selects an epoch of at least one. A fresh seed-42 model then refits the complete fitting window for that epoch count. TCN uses cuDNN algorithm benchmarking; forest inference uses serial tree accumulation for reproducible checkpoint replay.

Baseline scope and fixed references. The library comprises seven statistical rules, six linear/robust methods, seven neighborhood/tree methods, and six neural architectures. Task adaptation uses the shared causal features and trend-residual target described above. Project-specific multi-model residual pipelines and the HGB+MLP ensemble are not counted as standard algorithms. For each crop, the lowest thirteen-year mean annual RMSE identifies a descriptive reference within this library; its identity remains fixed across years, seeds, and cutoffs. Its descriptive role is defined in Section 5.1.

B.6.2 Direct seasonal inputs and fitting

All six direct seasonal models use the same complete-quality input. Each active slot contains 24 metadata fields (calendar 5, current-product quality/support 6, weather validity 13), thirteen standardized weather variables, and an unobserved-window indicator. Each selected product adds seven channels: previous-year value, anomaly, validity, and quality; and observed-window value, anomaly, and validity. Static features comprise history/context 20, crop coverage 1, and 24 trajectory-summary features per product. Single-product inputs have shape 12×45 plus 45 static features; two-product inputs have shape 12×52 plus 69 static features. Ridge, Random Forest and LightGBM receive these inputs flattened to 585 or 693 dimensions.

Future values are masked before summaries or availability are computed. All six future product-quality fields are replaced with historical grid/month support estimated from the relevant fitting window. Training-only historical mean defines anomalies, and missing/inactive entries follow the same masked encoder for all six models. Unselected products' previous values or previous quality do not enter; current observed-window support from all products remains shared with the reference seasonal interface.

Direct GRU and Transformer each have two layers of width 128; the Transformer has four attention heads. The shared 128/64-wide output head uses dropout 0.1. Optimization uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 2,048, gradient clipping at 1, at most 40 epochs, and patience 6. Training years receive equal total sample weight. LightGBM compares two preregistered capacities: 15 leaves/minimum leaf 200/L2 penalty 10, or 31 leaves/minimum leaf 40/L2 penalty 1. Both use learning rate 0.03, row/feature fractions 0.9, at most 1,200 trees, and 60-round early stopping. Mean annual internal validation RMSE selects the candidate and duration. Each chosen configuration is refitted through the block cutoff.

Direct Ridge selects $\alpha \in \{1, 10, 100, 1000, 10000\}$ using mean annual internal-validation RMSE; feature scaling is fitted anew in each fitting window. Random Forest retains 128 trees, depth 20, minimum leaf size 8, feature fraction 0.75 and bootstrap sampling, with no new capacity search. Both optimize year-weighted residual targets.

The CNN–RNN control adapts seasonal convolution and historical recurrence to the common inputs, following the modeling pattern of Khaki et al. (2020). Two kernel-3 Conv1D layers of widths 64 and 128 encode the seasonal sequence, with ReLU, dropout 0.1, per-layer activity masking and masked mean pooling. A one-layer, 64-wide LSTM encodes five historical yields and validity indicators. The seasonal embedding, history embedding and static features feed a 128/64-wide output MLP. Optimization and year weighting match the direct GRU; cuDNN benchmarking is enabled. This

is an input-matched adaptation, not a reproduction of the original architecture or its additional soil, management and multi-year weather inputs.

There are 216 final direct models: four crops, three fitting cutoffs, three issue fractions, and six model families. The historical library contains 312 method–crop–block configurations, including statistical rules; this count is not 312 neural trainings. All models are kept in the comparisons, rather than selecting a different subset at each crop or cutoff.

B.6.3 PROVENANCE AND RELEASE

Each result stores fitting/selection windows, sample identities, normalization, fitted capacity, parameters, and input/code hashes. Predictions share the same crop–grid–year targets. The world-model pipeline and all learned baselines use the common preceding-validation and full-window refitting protocol in Table 9. Earlier shorter-window results remain separate.

A released, relocatable CPU inference bundle covers 96 fixed grid–year samples from 2006–2008, four crop recipes, and 10/30/50% unobserved windows. It includes original historical experts, BIID weights, trees, and observed-window-only inputs. Nine checks pass in a fresh CPU-only environment with a Python-level guard against reads from the original workspace. Archived-trajectory features and yields reproduce exactly. Full CPU-FP32 inference differs from the original mixed-precision predictions by at most 0.0132 t/ha on these samples; reported scores are unchanged. This checks frozen inference, not complete preprocessing or full retraining. The processed-input release and source registry retain provider-specific terms; the bundle does not substitute for the complete set of benchmark checkpoints.

B.7 IMPLEMENTATION AND REPRODUCIBILITY

The implementation separates physical input preparation, state fitting, historical experts, yield readouts, and evaluation. Their schemas, temporal boundaries, and feature definitions are specified in Appendix B; repetitions and regional refits are specified in Appendices C.3.1, C.4, and C.5. The accompanying anonymous release specifies input identities, configurations, dependencies, selected recipes, and reference outputs for each component.

Independent-environment component reconstruction. Using exported training inputs and fresh dependency environments, we reconstruct the 45 LightGBM branches, 36 historical MLPs, nine soybean TabM components, and 45 vegetation models used by the 36 crop–block–seed systems. Reconstruction preserves the training durations and tree counts selected on the preceding validation years, including the six zero-epoch TabM components specified in Appendix C.3.1. Reconstructed MLP predictions supply the TabM residual targets. Tree structures, leaf values, and neural parameters match the reference components exactly. This reconstruction verifies the selected components, not the original search process; selection provenance requires the corresponding candidate scores and upstream-checkpoint records.

Complete evaluation with reconstructed components. The reconstructed components are evaluated on the original partitions, preserving mixed-precision and batch settings. The complete-observation reference and two completion rules at three cutoffs reproduce all 252 output conditions and 1,092 annual scores. Fitting and forward evaluation are isolated from reference-output files; those outputs are used only for subsequent numerical comparison. State inputs expose the observed window and mask withheld vegetation.

Numerical precision. Matching the original CUDA mixed-precision settings reproduces the saved outputs exactly. A portable CPU forward check on 96 grid–year examples has a maximum prediction difference of 0.01317 t/ha relative to CUDA. Numerical comparisons therefore retain the precision and batching settings of the corresponding reference computation.

Public artifacts and reproduction boundary. The anonymous repository releases raw-product preprocessing entry points, model and evaluation code, protocol files, selected fitting recipes, and machine-readable annual scores used to rebuild the main tables. Its linked processed-data release contains the selected crop-active yield, ERA5-Land, LAI, NDVI, GPP, and SEAS5 numeric inputs

(5.88 GB compressed; 47.68 GB extracted), with manifests and checksums. A four-crop sample additionally provides frozen historical, state, and yield-head weights for complete BIID-to-yield inference at four cutoffs.

The release does not include every exploratory cache, the complete candidate score history, or all checkpoints used across the full benchmark. Although the original workflow and training sources are provided, a fresh-clone rerun of every search and training job is not claimed as a turnkey reproduction. Raw provider archives are not redistributed; users obtain them under their version-specific terms and run the released preprocessing entry points. Independent-environment reconstruction above is an author-side verification; the public reviewer-facing checks comprise the executable inference sample, processed-input integrity checks, released annual scores, and table rebuilding.

C Additional Experimental Results

This appendix reports the full baseline comparisons, spatial analyses, seed repetitions, independent regional refits, and seasonal-weather tests. The original experiment populations and numerical results are retained.

C.1 Complete Thirteen-Year Results

This section extends the main-text comparisons across observation cutoffs and spatial support. All exhibits use the same thirteen evaluation years; relative reductions compare equally weighted mean annual RMSEs.

Complete baseline results. Table 11 reports the historical library and PhenoBIID across 10–100% unobserved windows. Table 12 retains the broad seed-42 direct seasonal comparison at a 10% unobserved window. The main-text Table 1 consolidates the repeated strong direct-model comparison at 10%, 30%, 50%, and 70%. Figure 12 resolves the historical-reference comparison by year.



Figure 12: Annual RMSE reduction versus the crop-specific historical reference at a 10% unobserved window, seed 42.

Full observation sweep. Figure 13 compares BIID and historical-mean completion across the full range of unobserved windows against the same historical reference. The main text emphasizes repeated-seed comparisons at 10%, 30%, 50%, and 70%.

For maize at a 10% unobserved window, the world-model pipeline reduces the mean annual RMSE of Direct LightGBM by 0.39%, with lower error in 7/13 years. This comparison follows the pre-specified primary metric, the equally weighted mean of annual spatial RMSE, used for every method and crop.

Figure 15 jointly evaluates the named historical reference and all six direct families using 1,000 paired spatial-block draws. Each draw retains all years within its sampled blocks. The improvement over the strongest direct method has positive intervals at both block scales for soybean and wheat, while maize and rice retain positive point estimates. Historical-reference comparisons are likewise positive across crops at the point-estimate level. The plot complements the seed repetitions by showing where the geographic evidence is strongest under the fitted models.

Figure 14 summarizes all six direct families at the three cutoffs. PhenoBIID has lower mean annual RMSE in each of the 72 crop–model–cutoff comparisons for seed 42. The tree ensembles provide the closest competition, with the smallest margins against LightGBM for maize and wheat. The

Table 11: Mean annual RMSE (t/ha) against 26 historical algorithms and PHENOBIID at 10–100% unobserved windows, seed 42 of the pipeline in Table 1. The 10%, 30%, and 50% PHENOBIID rows are the BIID + observed-feedback scores in Table 18 and the Full-input scores in Table 21. Lowest PHENOBIID window in bold; second-lowest and best historical method underlined.

Method	Maize	Rice	Soybean	Wheat
Statistical forecasting				
Global mean	4.2376	3.5524	1.1539	3.0617
Grid historical mean yield	1.5632	1.3675	0.4945	1.0154
Previous year	1.0591	0.9331	0.3832	0.9890
3-year moving mean	0.9321	0.8054	0.3328	0.8171
5-year moving mean	0.9321	0.8223	<u>0.3250</u>	0.8076
Exponential smoothing	0.9296	0.8040	0.3297	0.8169
Per-grid linear trend	0.9682	0.8410	0.3348	0.8490
Linear and robust regression				
Linear regression	0.9346	0.8182	0.3312	0.8980
Ridge regression	0.9345	0.8179	0.3310	0.8958
Lasso regression	0.9290	0.8156	0.3300	0.8450
Elastic Net	0.9294	0.8161	0.3299	0.8564
Huber regression	0.9437	0.8267	0.3312	0.8416
Linear SVR	0.9346	0.8182	0.3312	0.8974
Neighborhood and tree models				
Approximate KNN	0.9765	0.8374	0.3915	0.8693
Decision tree	1.0121	0.9213	0.4524	0.9146
Random forest	0.9151	0.8247	0.3961	0.8180
Extra Trees	0.9027	0.8255	0.3940	<u>0.8063</u>
Histogram gradient boosting	0.9016	0.8107	0.3808	0.8232
LightGBM	0.9025	0.8162	0.3377	0.8208
XGBoost	0.9082	0.8089	0.3376	0.8281
Neural predictors				
MLP	0.8986	<u>0.7852</u>	0.3268	0.8142
RNN	0.9088	0.7985	0.3347	0.8154
LSTM	0.9041	0.7878	0.3308	0.8140
TCN	<u>0.8985</u>	0.7853	0.3331	0.8145
GRU	0.9085	0.7939	0.3333	0.8321
Transformer	0.9105	0.7868	0.3334	0.8314
PHENOBIID (10%)	**0.8409**	**0.7532**	**0.2950**	**0.7642**
PHENOBIID (20%)	<u>0.8419</u>	<u>0.7539</u>	<u>0.2955</u>	<u>0.7648</u>
PHENOBIID (30%)	0.8429	0.7568	0.2982	0.7660
PHENOBIID (40%)	0.8452	0.7587	0.3002	0.7668
PHENOBIID (50%)	0.8474	0.7601	0.3022	0.7720
PHENOBIID (60%)	0.8510	0.7622	0.3047	0.7769
PHENOBIID (70%)	0.8561	0.7662	0.3083	0.7803
PHENOBIID (80%)	0.8571	0.7675	0.3092	0.7813
PHENOBIID (90%)	0.8584	0.7708	0.3132	0.7846
PHENOBIID (100%)	0.8583	0.7713	0.3132	0.7849

Table 12: Direct seasonal comparison at a 10% unobserved window, seed 42 of the complete pipeline. PHENOBIID matches Table 11. Mean annual RMSE (t/ha). Lowest in bold; second-lowest underlined.

Method	Maize	Rice	Soybean	Wheat
Ridge	0.9793	0.9062	0.3489	0.8439
Random Forest	0.8498	<u>0.7705</u>	<u>0.3126</u>	0.8078
LightGBM	<u>0.8442</u>	0.7791	<u>0.3126</u>	<u>0.7935</u>
GRU	0.8699	0.8088	0.3446	0.8297
Transformer	0.8993	0.8273	0.3521	0.8413
CNN–RNN (adapted)	0.8704	0.8049	0.3377	0.8315
PHENOBIID	**0.8409**	**0.7532**	**0.2950**	**0.7642**

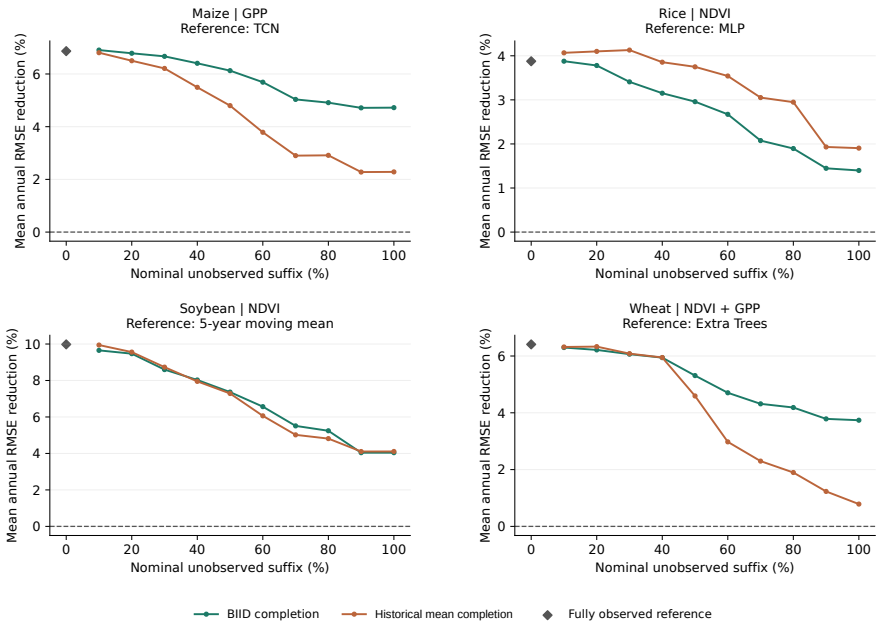

Figure 13: Mean annual RMSE reduction versus the historical reference across unobserved windows, seed 42. Positive values favor the named completion; $r = 0$ is fully observed.

Table 13: Internal historical anchors of PHENOBIID, distinct from the external library in Table 11.

Crop	Internal historical anchor	Seasonal correction
Maize	Causal trend and historical MLP	Two branches, equal mean
Rice	Historical MLP	One LightGBM branch
Soybean	Historical MLP + TabM residual	One LightGBM branch
Wheat	Historical MLP	One LightGBM branch

common scale makes this distinction visible alongside the larger improvements over linear and neural predictors.

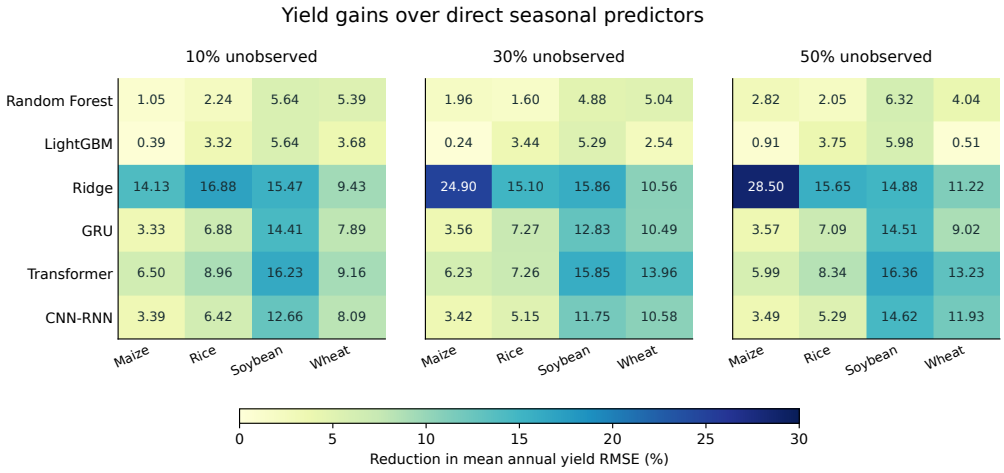

Figure 14: RMSE reduction (%) versus six direct predictors, seed 42. Panels fix the unobserved fraction. Positive values favor PhenoBIID.

The signal sweep reveals crop-specific and cutoff-dependent value. GPP is the strongest route for maize through the sweep, while NDVI is competitive for rice and soybean with longer observed windows. Combining NDVI and GPP is effective for wheat. These comparisons motivate evaluating product choice together with observation availability rather than assuming a single vegetation signal serves every crop equally well.

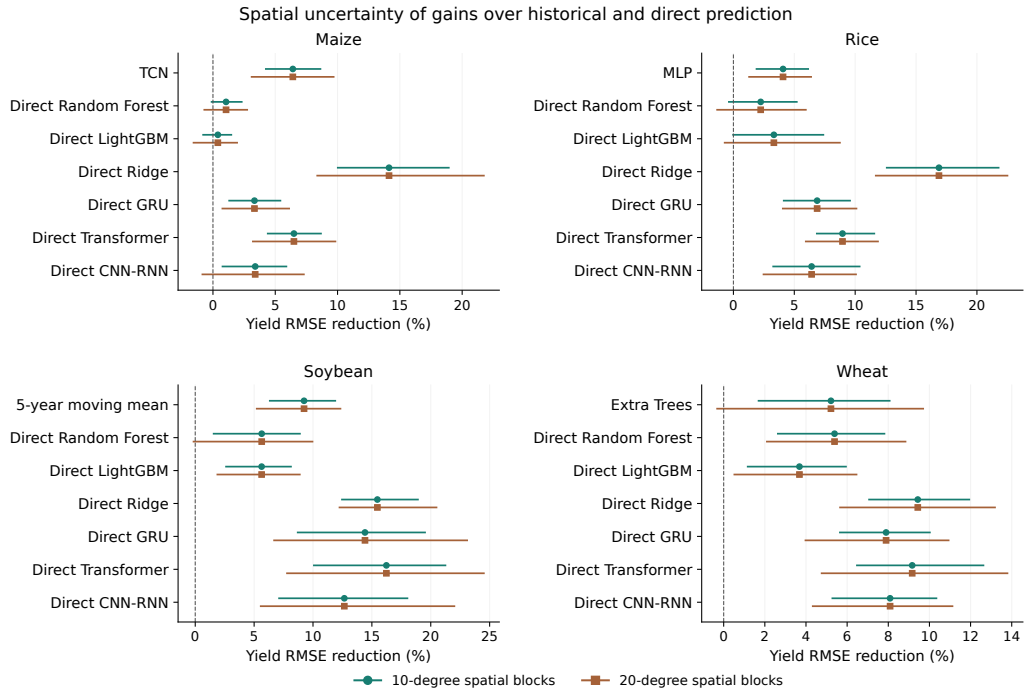

Figure 15: Paired 95% spatial-block intervals for RMSE reductions at 10%, seed 42. Positive values favor PhenoBIID.

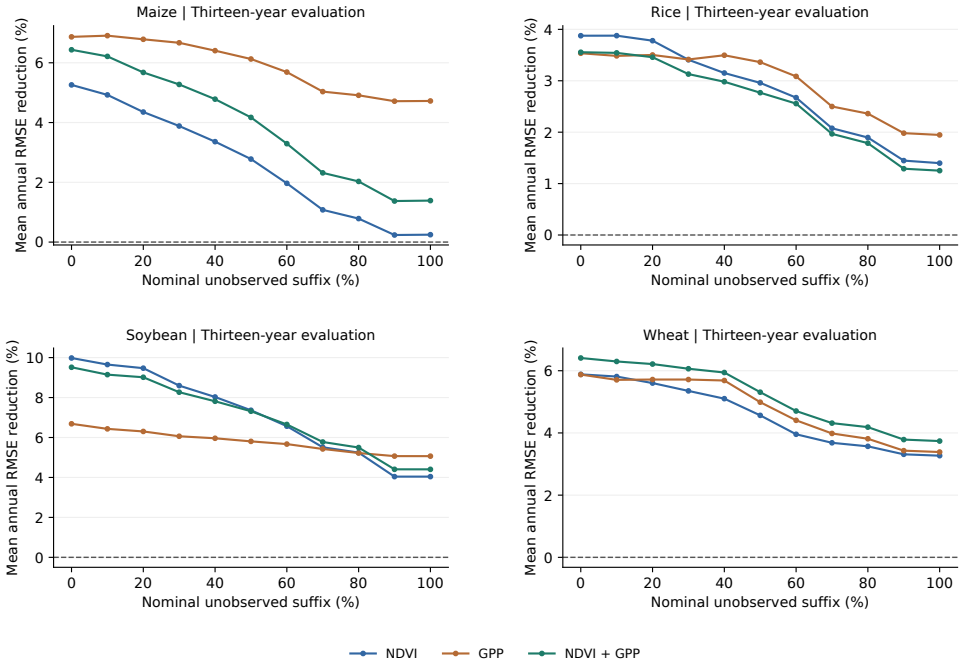

Figure 16: NDVI, GPP, and NDVI+GPP on the same thirteen-year cohort, relative to the historical reference in Table 11.

C.2 Spatial Yield Assessment

Figure 17 extends Figure 3 to prediction and signed error at a 10% unobserved window over all thirteen evaluation years.

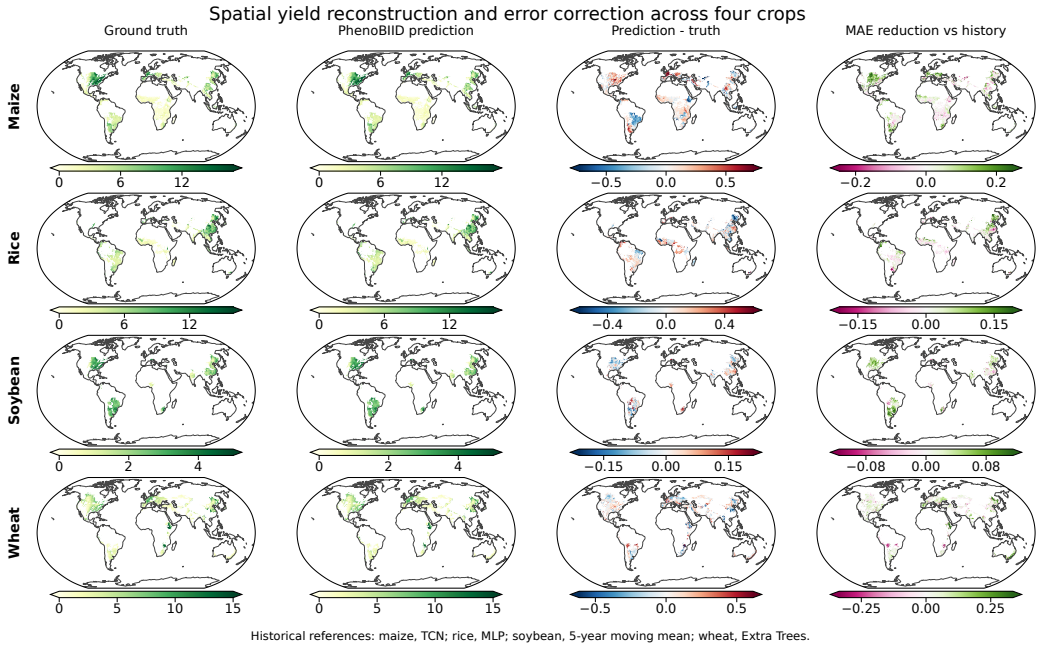

Figure 17: Spatial assessment at 10%, seed 42. Columns: truth, prediction, signed error, and paired absolute-error reduction (t/ha).

Each grid cell averages its available paired evaluation records; missing cells are unscored, not zero-yield observations. The named historical references are fixed as in Table 11. The last panel averages the per-record absolute-error difference, $|\widehat{y}^{\text{history}} - y| - |\widehat{y}^{\text{world}} - y|$, so positive values indicate improvement. Unlike the mean signed-error panels, it does not cancel under- and overprediction across years. The paired maps localize where the complete seasonal pipeline improves on historical prediction. Within each crop, truth and prediction share a scale. The signed error and absolute-error reduction use their own symmetric scales. Display limits retain the original true-yield 99th percentile and error 95th percentiles; colorbar extensions indicate saturated values. The four-row layout exposes the crop-specific spatial pattern while preserving all eligible grid cells.

C.3 Robustness and Completion Analysis

C.3.1 Complete-pipeline repetitions

We assess the stability of yield gains across random initializations by repeating the complete pipeline with seeds 42, 45, and 48. Crop-specific vegetation recipes, temporal partitions, and component training settings are held fixed across seeds. State models, historical components, and yield readouts are independently initialized and fitted according to their registered procedures. Each seed is evaluated separately on the same thirteen-year cohort against the fixed named historical references in Table 11; predictions are not ensembled across seeds.

For each seed, we first average annual yield RMSE over the evaluation years. Table 14 reports the mean and sample standard deviation of this metric across seeds, together with the corresponding relative reduction from the historical reference. Standard deviations describe initialization variability, whereas the spatial-block intervals in Appendix C.1 assess geographic uncertainty.

Yield gains persist across initializations. All three seeds outperform the historical reference in mean annual RMSE for every crop, with improvements observed in a majority of evaluation years. Similar robustness is observed under longer unobserved window settings. Soybean exhibits greater

Table 14: Three-seed RMSE and historical-reference gains at a 10% unobserved window. Parentheses are sample SD; positive years count evaluation years won.

Crop	RMSE, t/ha (SD)	Gain, % (SD)	Positive years
Maize	0.8440 (0.0031)	6.06 (0.35)	10–11/13
Rice	0.7545 (0.0013)	3.91 (0.16)	11–12/13
Soybean	0.2992 (0.0037)	7.94 (1.13)	12/13
Wheat	0.7639 (0.0007)	5.26 (0.09)	10/13

seed-to-seed variation, but its mean-RMSE advantage remains positive in every repetition. These results indicate that the complete-pipeline gains are not confined to a single initialization.

Predictive completion gains at longer horizons. We further evaluate the contribution of predictive vegetation completion by comparing BIID completion with historical-mean vegetation completion. In this comparison, the observed window, weather inputs, historical yield features, and yield-readout weights are kept identical, so only the completion rule differs. Both alternatives provide a completed vegetation trajectory; historical-mean completion is not a no-trajectory control.

Figure 4 in the main text compares the two rules across seeds. At 50%, mean completion gains are 1.38%, 0.56%, and 0.71% for maize, soybean, and wheat. At 70%, the corresponding reductions reach 2.38%, 0.88%, and 2.64%. Across these crops, BIID completion provides its clearest gains at longer unobserved windows, showing that predictive trajectories become increasingly valuable as less of the growing season is observed.

These results support distinguishing the shared use of completed vegetation trajectories from the additional information supplied by predictive completion beyond historical seasonal patterns. They also highlight longer-horizon prediction as the setting in which learned vegetation evolution contributes most clearly to downstream yield assessment.

C.4 Paired Strong Direct-Model Repetitions

We repeat the two strongest direct seasonal tree families, Random Forest and LightGBM, with seeds 42, 45 and 48 at 10%, 30%, 50%, and 70% unobserved windows. They use the same thirteen years and crop–grid–year cohort as the corresponding world-model seed, with the registered input encoding, normalization and training protocol unchanged. The four crops, three temporal blocks, three seeds and two direct families produce 72 fitted configurations on sample identities and labels matched to the world model.

The three-seed scores appear in Table 1. The broader seed-42 family comparison is reported separately in Table 12.

For each seed we first average annual spatial RMSE over thirteen years; Table 1 reports the mean and sample SD of these scores, not the performance of an ensemble. The gain reported in the main text is $100(1 - \mathrm{Mean}_s\, \overline{E}_c^{\mathrm{world},s} / \mathrm{Mean}_s\, \overline{E}_c^{\mathrm{direct},s})$. The descriptive stronger-direct reference is LightGBM for maize and wheat, and Random Forest for rice and soybean. It is held fixed across seeds and years; both families are reported in Table 1.

The world model has lower mean error than both direct families for all four crops. Relative to the lower-mean direct method, the reductions are 0.22%, 1.99%, 4.18% and 3.13% for maize, rice, soybean and wheat. Rice, soybean and wheat improve in all three paired seeds, and maize retains a positive three-seed mean. Across seeds, positive-year counts against these references are 6–7/13, 7–9/13, 6–7/13 and 10–11/13, respectively. These results show gains across all four crops, with particularly consistent seed-level improvements for rice, soybean and wheat.

The routes use different encoders and historical anchors, so these results characterize complete-method performance. Same-head completion experiments isolate the state contribution. The following cutoff analysis examines whether the direct-model advantage persists as observations become scarcer.

C.4.1 Paired comparisons across observation cutoffs

The 10%, 30% and 50% comparison contains 216 direct-model fits, pairing both direct families with the world model across four crops, three temporal blocks and three seeds at each cutoff. The 70%

extension adds 72 matched direct-model fits under the same protocol; its RMSE summary appears in Table 1.

Table 15: Paired RMSE reductions (%) versus each direct family. Positive years are ranges across seeds.

Crop	Direct reference	Unobs. (%)	Gain (%)	Positive seeds	Positive years
Maize	LightGBM	10	+0.22	2/3	6–7/13
Maize	LightGBM	30	+0.12	2/3	5–6/13
Maize	LightGBM	50	+0.47	3/3	4–5/13
Maize	Random Forest	10	+0.81	3/3	9–10/13
Maize	Random Forest	30	+1.79	3/3	9–10/13
Maize	Random Forest	50	+2.31	3/3	8–9/13
Rice	LightGBM	10	+3.59	3/3	10–10/13
Rice	LightGBM	30	+3.18	3/3	9–10/13
Rice	LightGBM	50	+3.77	3/3	9–10/13
Rice	Random Forest	10	+1.99	3/3	7–9/13
Rice	Random Forest	30	+1.51	3/3	7–8/13
Rice	Random Forest	50	+1.79	3/3	7–8/13
Soybean	LightGBM	10	+4.40	3/3	7–9/13
Soybean	LightGBM	30	+3.92	3/3	7–9/13
Soybean	LightGBM	50	+4.74	3/3	8–9/13
Soybean	Random Forest	10	+4.18	3/3	6–7/13
Soybean	Random Forest	30	+3.53	3/3	6–9/13
Soybean	Random Forest	50	+5.21	3/3	7–10/13
Wheat	LightGBM	10	+3.13	3/3	10–11/13
Wheat	LightGBM	30	+2.92	3/3	10–10/13
Wheat	LightGBM	50	+0.97	3/3	10–10/13
Wheat	Random Forest	10	+5.60	3/3	12–12/13
Wheat	Random Forest	30	+4.94	3/3	10–11/13
Wheat	Random Forest	50	+4.19	3/3	12–12/13

All 24 crop–cutoff–reference contrasts have positive seed-mean gains (Table 15), and 22 improve in all three seeds. At 50%, maize improves against LightGBM in mean error for every seed, while rice retains gains of 1.51–1.99% over Random Forest across the evaluated cutoffs. The paired table reports improvement magnitude, seed agreement and positive-year counts as complementary measures of robustness. At 70%, all eight crop–reference contrasts favor the world-model pipeline in all three seeds. Mean gains are 1.21–5.50% across direct families, with positive-year counts ranging from 5/13 to 12/13.

C.5 Independent Regional Statistical Labels

Question and scope. We test whether the conditional seasonal interface also supports yield prediction with regional statistical labels rather than GDHY's reconstructed grid yields. We refit the models on CY-Bench labels (Kallenberg et al., 2026) under the same conditional setting, using complete actual ERA5-Land as a supplied meteorological condition. The signals are GPP for maize and NDVI+GPP for wheat; rice and soybean are not evaluated here.

Cohort and alignment. The interface covers 509 administrative regions in Germany, France and Poland, intersecting 742 coarse cells through 2,449 nonempty overlaps. Monthly products are aggregated with equal-area polygon overlaps and valid product support, not crop-area weights. WorldCereal phase-I calendars define up to twelve packed seasonal slots; explicit source year–month indices retain cross-year seasons. Monthly boundary handling does not provide day-exact phenology. Germany's winter-wheat definition differs from the soft-wheat series in France and Poland.

The three frozen-weight temporal blocks match Table 9; only regions represented before each fitting cutoff are evaluated. Germany and France each contribute thirteen years, whereas Poland contributes 2010–2016 only. The fixed cohort has 9,657 region–year labels: maize contributes 2,606, 1,222 and 118 in Germany, France and Poland; wheat contributes 4,467, 1,126 and 118. Each route and seed is paired on these same labels. Scores first average annual spatial RMSE within a country, then summarize seeds 42, 45 and 48; countries are not pooled into the global main score.

Models and fitting. All regional model weights are newly fitted under preceding validation. The history comparison retains five causal formulas and Ridge, Random Forest and LightGBM. A separate history-selection strategy chooses among the three learned families using only preceding validation: LightGBM is selected for maize and Ridge for wheat. The comparison reports both the eight algorithms and this validation-selected strategy. Fixed formulas and deterministic Ridge copies are not independent random repetitions.

State models use the existing 128-dimensional, twelve-token BIID transition, with two interaction layers, observation feedback and vegetation supervision; they never receive the target yield as input. State fitting selects a positive training duration, with a maximum of thirty epochs. The selected history predictor is frozen. A residual LightGBM head is fitted on complete observed trajectories and then evaluated with the observed window plus either a BIID-predicted unobserved window or training-period historical mean. This is staged training, not end-to-end yield supervision. Direct Random Forest and LightGBM use the same regional structured input and observed window without predicting vegetation, separately fitted at each cutoff. Single-product and two-product terminal encodings contain 674 and 770 features, respectively; these regional calendar and support encodings differ from the global 501/537-feature interface.

The fixed experiment contains 27 state fits, 54 learned-history fits, 108 direct fits and 18 terminal heads. Two pre-registered LightGBM configurations are selected using preceding validation; Random Forest uses 128 trees, depth 20, minimum leaf size 8 and feature fraction 0.75. All routes use the same cutoff-specific observation mask; withheld vegetation values and support are excluded from the issue-time inputs.

Table 16: Mean annual RMSE (t/ha) on independent regional labels, averaged over three seeds (sample SD). Lowest in bold; second-lowest underlined. Results use a 10% unobserved window.

Method	Maize DE	Maize FR	Maize PL	Wheat DE	Wheat FR	Wheat PL
Latest available yield	1.598 (0.000)	1.676 (0.000)	1.225 (0.000)	0.991 (0.000)	0.980 (0.000)	0.443 (0.000)
Past 3-yield mean	1.233 (0.000)	<u>1.413</u> (0.000)	1.026 (0.000)	0.881 (0.000)	0.796 (0.000)	0.420 (0.000)
Past 5-yield mean	1.245 (0.000)	1.458 (0.000)	0.968 (0.000)	0.815 (0.000)	0.740 (0.000)	0.416 (0.000)
Past yield mean	1.930 (0.000)	1.728 (0.000)	1.112 (0.000)	1.225 (0.000)	0.736 (0.000)	0.549 (0.000)
Past yield trend	1.279 (0.000)	1.557 (0.000)	1.001 (0.000)	0.909 (0.000)	0.774 (0.000)	0.400 (0.000)
Ridge (history)	1.206 (0.000)	1.538 (0.000)	0.971 (0.000)	0.821 (0.000)	0.728 (0.000)	0.386 (0.000)
Random Forest (history)	1.235 (0.002)	1.495 (0.004)	0.890 (0.005)	0.862 (0.002)	0.760 (0.004)	<u>0.384</u> (0.001)
LightGBM (history)	1.237 (0.001)	1.492 (0.006)	0.927 (0.021)	0.885 (0.002)	0.751 (0.003)	0.390 (0.001)
Validation-selected history	1.237 (0.001)	1.492 (0.006)	0.927 (0.021)	0.821 (0.000)	0.728 (0.000)	0.386 (0.000)
Random Forest (direct)	<u>1.093</u> (0.007)	1.489 (0.003)	0.743 (0.009)	0.817 (0.006)	0.727 (0.002)	0.437 (0.011)
LightGBM (direct)	1.113 (0.001)	1.456 (0.005)	<u>0.685</u> (0.006)	0.822 (0.017)	0.721 (0.012)	0.389 (0.007)
Historical-mean completion	1.143 (0.018)	1.422 (0.003)	0.723 (0.040)	<u>0.805</u> (0.002)	<u>0.702</u> (0.000)	**0.368** (0.008)
PHENOBIID	**1.081** (0.017)	**1.406** (0.002)	**0.676** (0.041)	**0.798** (0.002)	**0.695** (0.001)	0.388 (0.010)
Full observed diagnostic	1.141 (0.018)	1.424 (0.002)	0.723 (0.040)	0.808 (0.002)	0.712 (0.001)	0.387 (0.007)

Results. At a 10% unobserved window, PHENOBIID has the lowest seed-mean RMSE in five of the six crop–country groups (Table 16). Polish wheat is the exception. Gains against the validation-selected history, averaged across paired seeds, are 12.61%, 5.76% and 27.08% for maize in Germany, France and Poland, and 2.80% and 4.53% for wheat in Germany and France. The largest improvement occurs for Polish maize, complementing the gains observed in the larger German and French cohorts.

In these five groups, PHENOBIID also improves on both direct regression families. The magnitude of improvement varies across regions, reflecting differences in crop dynamics, available sample size, and regional predictability.

Cutoffs and mechanism. Table 17 further examines the effect of prediction stages using three unobserved window lengths and four paired references. In five of the six crop–country groups, PHENOBIID keeps positive mean gains against the validation-selected history, both direct families, and historical-mean completion at every evaluated cutoff. Polish wheat does not. Both crops in Germany and France, and maize in Poland, therefore favor PHENOBIID across these references.

BIID and historical-mean completion share the same terminal head, enabling a controlled assessment of predicted vegetation dynamics. The comparison shows that BIID completion provides additional benefits over historical vegetation patterns in several regional settings, with maize exhibiting increasing relative advantages as longer unobserved windows are predicted across all three countries. The remaining variation across crops and regions reflects differences in crop-specific growth dynamics and predictability.

Table 17: Paired RMSE reductions (%) on independent regional labels. Positive values favor PHENOBIID.

Crop	Country	Unobs. (%)	History	Direct RF	Direct LGBM	Historical-Mean Completion
Maize	DE	10	+12.61	+1.10	+2.88	+5.42
Maize	DE	30	+12.20	+0.85	+2.40	+5.70
Maize	DE	50	+12.45	+1.05	+2.55	+6.35
Maize	FR	10	+5.76	+5.57	+3.43	+1.13
Maize	FR	30	+5.20	+5.90	+3.55	+1.35
Maize	FR	50	+4.70	+5.40	+3.90	+1.80
Maize	PL	10	+27.08	+9.02	+1.31	+6.50
Maize	PL	30	+25.40	+9.40	+1.10	+7.05
Maize	PL	50	+24.80	+9.20	+0.85	+7.60
Wheat	DE	10	+2.80	+2.33	+2.92	+0.87
Wheat	DE	30	+2.10	+1.80	+1.60	+0.55
Wheat	DE	50	+2.25	+1.95	+1.45	+0.70
Wheat	FR	10	+4.53	+4.40	+3.61	+1.00
Wheat	FR	30	+3.40	+4.10	+3.10	+0.40
Wheat	FR	50	+3.20	+3.60	+2.70	+0.55
Wheat	PL	10	-0.47	+11.26	+0.23	-5.27
Wheat	PL	30	-6.97	+5.94	-1.66	-4.62
Wheat	PL	50	-6.88	+4.98	-0.46	-4.39

Figure 18 complements the mean performance comparison by illustrating annual variations. State errors are evaluated only on hidden slots with valid supervision, separately for NDVI and GPP rather than averaging their different physical units. For wheat, the joint terminal prediction prevents attributing the final yield improvement to either vegetation product individually.

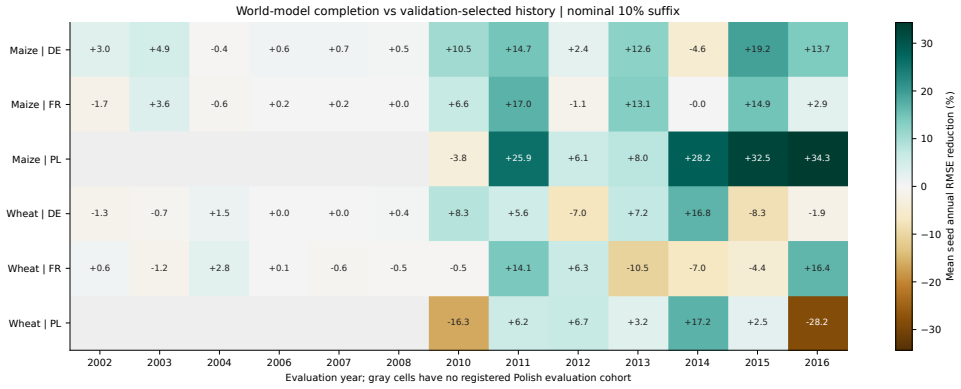

Figure 18: Regional annual RMSE reductions at 10%, averaged across seeds.

C.6 SEASONAL FORECAST WEATHER IN THE UNOBSERVED WINDOW

This appendix specifies the Section 5.5 substitution test. It uses the same crop–grid population and the same unobserved-window rule as Section 5.1 ($\lceil rM \rceil$ hidden active slots). This focused substitution test covers nine evaluation years, seed 42, and the 10%, 30%, and 50% windows, complementing the broader realized-weather evaluation in Table 1.

C.6.1 DATA AND YEARS

Observed vegetation, ERA5-Land weather, and yield labels are the CROPDYNAMICSBENCH records of Section 5.1. SEAS5 is the ECMWF System 51 monthly hindcast, ensemble-mean 2 m temperature, total precipitation, and surface solar radiation downwards, interpolated to the 0.5° crop grid. Evaluation years are 2002–2004, 2006–2008, and 2010–2012 (nine years; the 2010–2016 block is truncated), seed 42, at 10%, 30%, and 50% unobserved windows.

C.6.2 INFORMATION BOUNDARY

Observed-window weather remains ERA5-Land. Only hidden crop-active slots are replaced. The issue time is the start of the first hidden month; the previous calendar month's hindcast is used so a run issued during that month is not treated as available at its beginning. SEAS5 leads 1–6 are admitted; lead 1 is the initialization month, so the first hidden month is lead 2. Hidden slots beyond lead 6 use training-period grid–month climatology, so the reported score is SEAS5 where supported and climatology otherwise. Cover in Table 3 is the fraction of crop–grid–year samples whose entire unobserved window lies within those leads. Bias correction is fit on the corresponding training window. The remaining ten weather channels in hidden slots are set to training means. Direct Random Forest and LightGBM models are not retrained under SEAS5.

C.6.3 MODELS AND METRIC

State models are three-variable prefix-feedback fits. Yield heads are the original paper heads, not the Table 1 refits. The yield metric is the same annual spatial RMSE, averaged equally over the nine years. Wheat has median 11 crop-active months. At a 50% unobserved window most samples hide six months, which exceeds the usable SEAS5 horizon; 17% of wheat samples are fully covered. Table 3 reports the resulting hybrid full-sample score; Figure 8 uses the same sample set. The historical-weather completion is reported alongside SEAS5 to separate forecast substitution from the value of the observed vegetation prefix. vs ERA5, vs RF, and vs LGB are $100\big(\overline{E}_{\mathrm{SEAS5}}/\overline{E}_{\mathrm{ref}} - 1\big)$ from equal-year means of annual spatial RMSE; positive values denote higher RMSE than the named reference. vs ERA5 uses the same yield heads under realized weather. vs RF and vs LGB restrict the Table 1 Random Forest and LightGBM models to seed 42 and these nine years, still under realized weather; they are not retrained under SEAS5.

D ADDITIONAL ANALYSIS OF PHENOBIID

The following controls distinguish vegetation forecasting accuracy from its downstream yield value, and separate state-weather, terminal-input, and training-distribution effects.

D.1 STATE ARCHITECTURE AND COMPLETION CONTROLS

We hold the blockwise refitted historical experts, terminal yield heads, weather condition, observed window and sample identities fixed, changing only the rule completing the unobserved vegetation window. The BIID + observed-feedback yields match Table 11 at the same windows. Six rules are evaluated: BIID, GRU, BIID without observed feedback, training-period grid–month historical-mean vegetation completion, the previous trajectory, and persistence of the last visible observation. The last three use the fallbacks in Appendix D.5. Figure 19 focuses on BIID and the three simple completion rules; the GRU and feedback controls remain in Tables 18–20.

Matched GRU state model. The initial 128-dimensional state is a linear projection of the previous trajectory, validity, quality and five context fields. A 128-wide GRUCell receives each slot's thirteen weather variables, preceding-year value, validity and quality, the latest feedback value, and a twelve-dimensional position indicator. Its normalized 64-wide readout predicts a residual to the preceding-year value. Active slots update the recurrence; inactive slots leave it unchanged. At inference, a finite observed-window value replaces predicted scalar feedback, with the same issue mask as BIID. Only the observed window supplies observational feedback.

Both state families use AdamW with learning rate $3 \cdot 10^{-4}$, weight decay 10^{-4}, batch size 256, gradient clipping at 1, at most 30 epochs, and patience 5. Preceding-year validation selects a trained epoch, followed by a fresh full-window refit. Fifteen GRU fits cover nine NDVI and six GPP crop–block conditions. Architectures have different parameter counts but the same data, targets and optimization budget.

Feedback intervention. The no-observed-feedback control retains the trained BIID weights and autoregressively feeds predictions instead of observations in the observed window. The true observed window is still passed to the terminal head. Thus this intervention isolates observed-window as-

similation in the state recurrence; it is neither a no-recurrence model nor a retrained no-feedback ablation.

Table 18: Architecture and feedback controls, seed 42. Mean annual yield RMSE (t/ha). Lowest in bold; second-lowest underlined. BIID + observed feedback is the same PHENOBIID completion as Tables 11 and 21 at these windows. GRU and no-feedback change only the completion rule; no-feedback retains trained BIID weights and the observed window.

Unobserved	Completion	Maize	Rice	Soybean	Wheat
10%	BIID + observed feedback	**0.8409**	**0.7532**	**0.2950**	**0.7642**
10%	GRU + observed feedback	0.8430	0.7537	0.2952	0.7646
10%	BIID without observed feedback	0.8416	0.7549	0.2956	0.7644
30%	BIID + observed feedback	**0.8429**	**0.7568**	**0.2982**	0.7660
30%	GRU + observed feedback	0.8450	0.7612	0.2983	0.7669
30%	BIID without observed feedback	0.8435	0.7579	0.2983	**0.7659**
50%	BIID + observed feedback	**0.8474**	**0.7601**	**0.3022**	**0.7720**
50%	GRU + observed feedback	0.8493	0.7612	0.3025	0.7723
50%	BIID without observed feedback	0.8477	0.7608	0.3024	0.7722

Table 19: Mean annual hidden-slot RMSE for matched completion rules. Units are NDVI or GPP gC m^{-2} day^{-1}. Lowest in bold; second-lowest underlined. Averaging the BIID column with the crop–product–window weights of Table 2 recovers the Full BIID vegetation entries there. No obs. removes observed-window feedback.

Crop	State	Unobs.	BIID	GRU	No obs.	Hist. mean	Previous	Last
Maize	GPP	10%	**0.5232**	0.6219	0.6462	0.7370	0.6815	1.2565
Maize	GPP	30%	**0.5259**	0.6310	0.6359	0.7574	0.7007	1.4630
Maize	GPP	50%	**0.5253**	0.6253	0.6299	0.7597	0.7022	1.4963
Rice	NDVI	10%	**0.0365**	0.0461	0.0460	0.0476	0.0456	0.0980
Rice	NDVI	30%	**0.0380**	0.0460	0.0461	0.0465	0.0464	0.1103
Rice	NDVI	50%	**0.0393**	0.0472	0.0473	0.0471	0.0476	0.1222
Soybean	NDVI	10%	**0.0405**	0.0487	0.0490	0.0471	0.0532	0.0996
Soybean	NDVI	30%	**0.0411**	0.0493	0.0495	0.0481	0.0541	0.1153
Soybean	NDVI	50%	**0.0415**	0.0495	0.0494	0.0481	0.0538	0.1428
Wheat	NDVI	10%	**0.0391**	0.0556	0.0539	0.0549	0.0537	0.1031
Wheat	NDVI	30%	**0.0400**	0.0524	0.0516	0.0556	0.0536	0.1785
Wheat	NDVI	50%	**0.0404**	0.0520	0.0513	0.0554	0.0543	0.1481
Wheat	GPP	10%	**0.3326**	0.4864	0.5053	0.5704	0.4741	1.5556
Wheat	GPP	30%	**0.4171**	0.6130	0.6119	0.6249	0.5621	2.9107
Wheat	GPP	50%	**0.4427**	0.5971	0.5991	0.6570	0.5963	2.3537

Table 20: Anomaly correlation after removing training-period grid–month historical means. Highest in bold; second-highest underlined. Win columns count years with lower RMSE than historical mean.

Crop	State	Unobs.	BIID	GRU	No obs.	Previous	BIID wins	GRU wins
Maize	GPP	10%	**0.721**	0.623	0.608	0.448	13/13	12/13
Maize	GPP	30%	**0.713**	0.624	0.617	0.442	13/13	12/13
Maize	GPP	50%	**0.711**	0.629	0.623	0.436	13/13	12/13
Rice	NDVI	10%	**0.558**	0.428	0.432	0.361	11/13	5/13
Rice	NDVI	30%	**0.526**	0.423	0.421	0.350	10/13	3/13
Rice	NDVI	50%	**0.513**	0.416	0.415	0.345	10/13	3/13
Soybean	NDVI	10%	**0.555**	0.430	0.422	0.287	10/13	4/13
Soybean	NDVI	30%	**0.570**	0.445	0.437	0.302	11/13	5/13
Soybean	NDVI	50%	**0.561**	0.438	0.435	0.300	11/13	3/13
Wheat	GPP	10%	**0.782**	0.619	0.618	0.447	13/13	11/13
Wheat	GPP	30%	**0.760**	0.592	0.594	0.442	13/13	9/13
Wheat	GPP	50%	**0.757**	0.619	0.618	0.441	13/13	12/13
Wheat	NDVI	10%	**0.645**	0.427	0.443	0.325	13/13	3/13
Wheat	NDVI	30%	**0.640**	0.481	0.490	0.349	13/13	3/13
Wheat	NDVI	50%	**0.646**	0.499	0.506	0.346	13/13	3/13

BIID reduces state RMSE relative to GRU in all fifteen crop–product–issue conditions in Table 19 and also raises anomaly correlation. At a 10% unobserved window, BIID has the lowest yield RMSE on all four crops. At 30%, it remains lowest for maize, rice, and soybean. The feedback intervention

produces closely matched terminal scores, while the state metrics consistently favor BIID. These paired results support BIID's state representation and complement the longer-window yield analyses below.

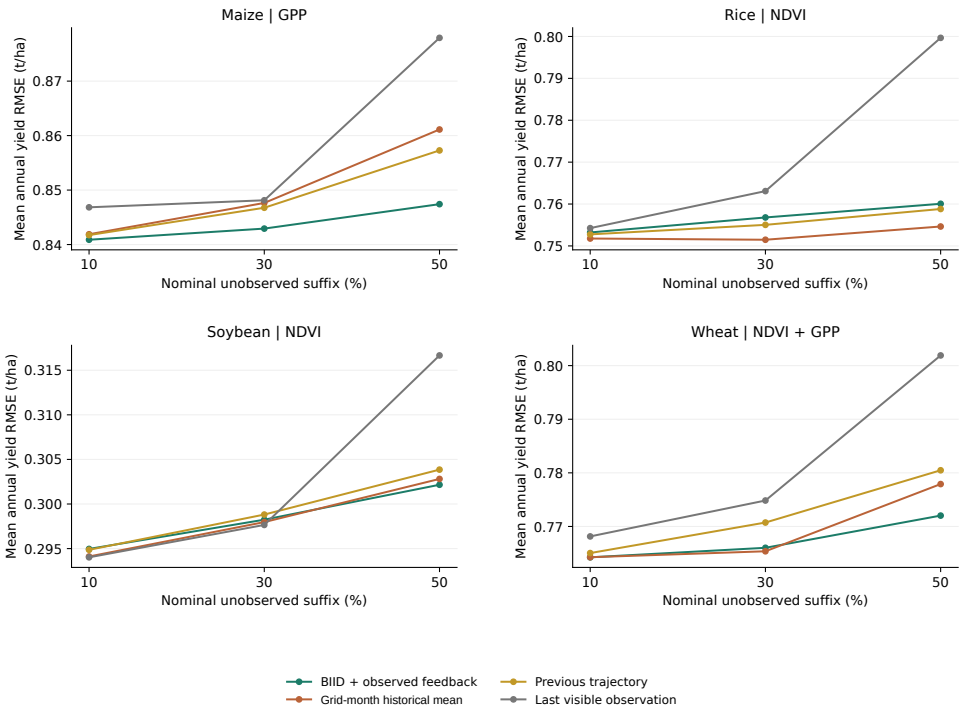

Figure 19: Yield RMSE of BIID versus simple completion rules under frozen heads, seed 42.

As the unobserved growth period lengthens, the value of forecasting vegetation trajectories for yield assessment becomes more evident across multiple crops. At the 50% cutoff, BIID yields lower error than all three simple rules for maize, soybean and wheat, demonstrating the downstream value of learned trajectory evolution when a substantial part of the season remains unobserved.

Figure 20 shows the 70% unobserved completion as predicted versus withheld vegetation, together with the resulting yield. Because crop-active seasons differ by location, the vegetation columns are relative positions along each grid's remaining window rather than a shared calendar.

D.2 Weather in the State-Transition Pathway

This ablation removes weather from state evolution while retaining it in the terminal yield input. Each scalar BIID model is trained again with all thirteen standardized weather channels fixed to zero in both fitting and inference. This means a constant normalized condition, not physically zero temperature or precipitation. The architecture, parameter count, previous vegetation, support fields, context and observed-window feedback are retained. Weather-token biases and contextual representations can still learn a background seasonal prior; the comparison therefore isolates the additional value of supplied weather variation.

Fifteen state models cover the five crop–product pairs at the three fitting cutoffs. Optimization and preceding-validation selection match Appendix D.1; every selected model has at least one trained epoch. The blockwise-refitted historical experts and yield heads are held fixed, and retain their original meteorological inputs. The ablation therefore tests the state-weather pathway, not removal of all weather information from the complete yield system. State and terminal scores use identical reference slots and grid–year identities.

Weather lowers mean annual state RMSE and increases anomaly correlation in all fifteen crop–product–issue conditions, with state RMSE reductions of 2.58–19.94% (Figure 6a). The yield effect is smaller: at a 50% unobserved window, retaining state weather lowers mean annual yield RMSE by 1.16% for maize, 0.66% for soybean and 0.57% for wheat. Thus the meteorological

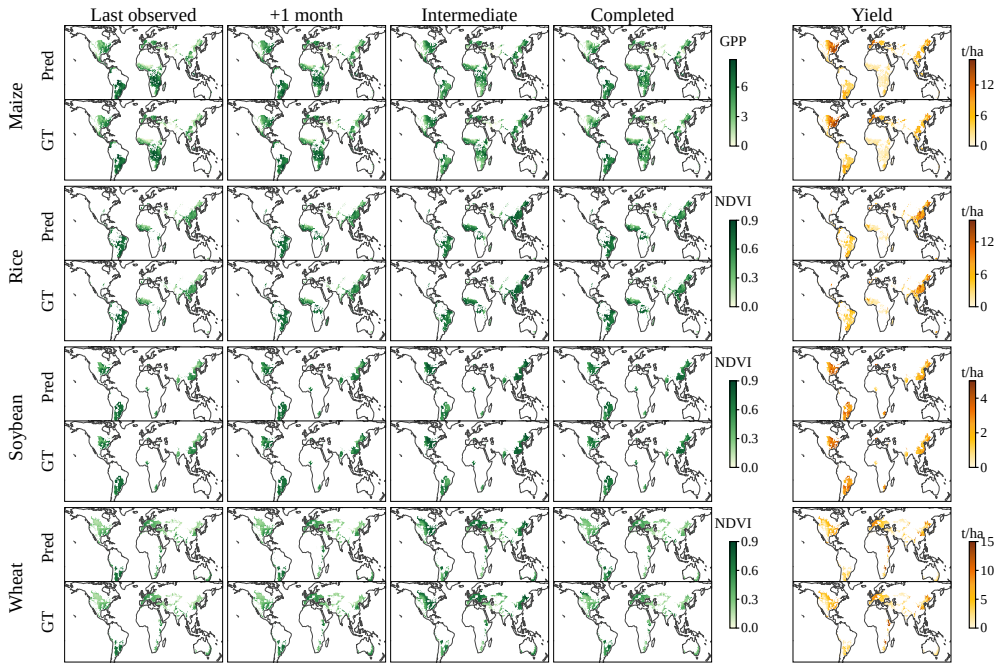

Figure 20: Predicted and withheld vegetation under a 70% unobserved window, plus the resulting yield. Crop-active seasons differ by location, so the vegetation columns are relative positions along each grid's remaining window (maize GPP; rice, soybean, wheat NDVI).

condition consistently improves state estimation and provides its clearest downstream benefit at longer unobserved windows. At shorter windows, the yield estimates remain stable because most of the vegetation trajectory is directly observed.

The paired state and yield effects are shown in the main-text Figure 6.

D.3 Terminal Weather and Product-Support Paths

Unlike the state-weather ablation, these controls keep the original weather-driven BIID predictions and historical experts fixed. We remove only terminal meteorological values and validity (312 columns), the six per-slot product-support fields (72 columns), or both. Calendar, crop coverage, historical features and completed-trajectory encodings remain. The input width stays at 501 or 537; removal is applied during both fitting and inference, and removed columns are verified to be unused by tree splits. Removing terminal support does not remove missingness effects in trajectory features or the quality information used to initialize the state.

Each branch retains its validation-selected tree parameters and tree count, seed, observed-trajectory training population, residual normalization and historical anchor. This fixed-capacity ablation retains the original fitted capacities and maize's two equally weighted branches. The four input conditions produce 48 crop–block–condition heads, which are frozen across three cutoffs and two completion rules, yielding 288 matched terminal evaluations.

At a 10% unobserved window, retaining direct terminal weather reduces mean annual RMSE by 1.50%, 0.94% and 0.81% for rice, soybean and wheat. Retaining terminal product support reduces error by 0.78%, 0.26%, 0.55% and 0.04%, respectively. The terminal weather path therefore contributes to three crops even when the same weather has already driven state evolution.

The Full-input RMSE values match the seed-42 PhenoBIID rows in Table 11 and the BIID + observed-feedback rows in Table 18. The bracketed gains in Table 21 isolate completion within each fixed head. Maize benefits from BIID under every terminal condition and cutoff. At 50%, soybean and wheat also benefit with or without terminal weather, showing that predictive vegetation retains value beyond direct meteorological access. Together, the two statistics show that terminal inputs and predictive completion provide complementary information for yield assessment.

Table 21: Yield RMSE (t/ha) and reduction (%) versus historical-mean completion under the same terminal head. Full input is the same seed-42 PHENOBIID completion as Tables 11 and 18 (BIID + observed feedback). Other rows remove terminal weather and/or support. Lowest RMSE in bold; second-lowest underlined.

Unobserved	Terminal input	Maize	Rice	Soybean	Wheat
10%	Full input	<u>0.8409</u> [+0.12]	**0.7532** [-0.19]	**0.2950** [-0.29]	**0.7642** [+0.00]
10%	No weather	**0.8401** [+0.12]	0.7647 [-0.34]	0.2978 [-0.23]	0.7705 [+0.03]
10%	No shared support	0.8475 [+0.13]	<u>0.7552</u> [-0.18]	<u>0.2966</u> [-0.23]	<u>0.7646</u> [-0.00]
10%	Neither path	0.8431 [+0.17]	0.7642 [-0.35]	0.2987 [-0.20]	0.7687 [+0.10]
30%	Full input	<u>0.8429</u> [+0.56]	**0.7568** [-0.71]	**0.2982** [-0.09]	**0.7660** [-0.08]
30%	No weather	**0.8427** [+0.60]	0.7676 [-0.85]	0.3011 [-0.04]	0.7723 [-0.08]
30%	No shared support	0.8498 [+0.52]	<u>0.7592</u> [-0.61]	<u>0.2989</u> [-0.04]	<u>0.7663</u> [-0.16]
30%	Neither path	0.8468 [+0.61]	0.7679 [-0.91]	0.3011 [+0.07]	0.7708 [-0.12]
50%	Full input	**0.8474** [+1.59]	**0.7601** [-0.72]	**0.3022** [+0.22]	**0.7720** [+0.76]
50%	No weather	<u>0.8479</u> [+1.59]	0.7710 [-0.96]	0.3051 [+0.26]	0.7782 [+1.09]
50%	No shared support	0.8541 [+1.45]	<u>0.7632</u> [-0.72]	<u>0.3026</u> [+0.18]	<u>0.7725</u> [+0.59]
50%	Neither path	0.8519 [+1.49]	0.7722 [-1.10]	0.3049 [+0.42]	0.7769 [+1.17]

D.4 INPUT VALUE IN DIRECT SEASONAL PREDICTION

D.4.1 INPUT REMOVAL

Figure 7 brings together input-group and product-level comparisons under direct LightGBM prediction. These controls test which inputs improve direct LightGBM yield regression, independently of the state-transition architecture. All conditions use the 10% issue point, identical grid–year samples, normalization rules, causal-trend anchor and year weighting. Each applies its input removal during both fitting and inference, with the same two LightGBM capacity candidates, preceding internal validation and full-window refitting as the complete-input baseline. Twelve complete-input models are reused and 36 removal models are fitted; no world-model state or historical expert is changed.

Without vegetation. We remove selected-product values, anomalies, summary features, validity and previous-year quality, as well as the six shared product-support fields. Historical/context features, crop coverage, the active calendar, issue mask, weather and weather validity remain. This tests vegetation information beyond the same meteorological and background inputs, without changing sample support.

Without weather. Both the thirteen meteorological values and their validity indicators are removed; vegetation and the remaining metadata are retained. Unlike Appendix D.2, this is a direct yield-model input ablation, not an intervention in a state forecaster.

Without shared product quality. Only the six common product-support fields are removed. The selected products' own validity and previous-year quality channels remain, together with their values and summaries. This separates the value of supplementary availability cues from that of the observed vegetation signals themselves. Calendar and crop coverage are not removed in this condition.

Removal sets the registered encoded fields to zero while preserving input dimensions. Saved tree split indices are checked to ensure that removed columns are unused. This is a retraining comparison, not post-hoc zeroing of the full model's input at evaluation time.

Retaining vegetation reduces mean annual yield RMSE by 3.92%, 3.35%, 4.10% and 4.16% for maize, rice, soybean and wheat, respectively, with lower annual error in 12, 10, 11 and 11 of thirteen years. This provides input-matched evidence that vegetation adds information beyond history, weather and calendar support even without a learned state-transition model. Weather and supplementary quality fields have smaller effects: retaining either provides additional aggregate gains for maize, rice, and soybean. Together with the four-crop vegetation result, this identifies vegetation as the most consistently informative added input and weather and quality fields as complementary signals whose value is crop dependent. The vegetation removal includes its availability fields, so it measures the whole vegetation input rather than isolating measured values from native validity indicators.

D.4.2 MATCHED LAI, NDVI AND GPP INPUTS

This complementary study compares LAI, NDVI, GPP and NDVI+GPP with one LightGBM architecture at a 10% unobserved window. All sixty crop–block–condition models are fitted anew, including a shared-background reference. The thirteen-year population, preceding validation, two capacity candidates, year weighting, causal-trend anchor and normalization rules are unchanged. The retained cohort already contains paired physical LAI trajectories; no signal-specific sample selection is applied.

The shared background contains history, location/time context, crop coverage, calendar, weather and the same six issue-masked product-support channels. Each vegetation condition adds its selected products' previous values and observed window, their deviations from historical means, summary statistics and native finite-value indicators. The background reference omits these added fields but retains shared availability information; it is not a history-only or satellite-metadata-free baseline.

The physical cache has no comparable native previous-LAI quality field. We therefore disable product-specific previous quality for *every* product, rather than treating LAI validity as measured quality. Shared background fields are unaffected. Single-product inputs retain 585 flattened features, and NDVI+GPP retains 693. The resulting models are distinct from the complete-quality direct models in Table 12. Their signal ranking is descriptive and does not reselect the world-model recipe or evaluate a learned LAI forecaster.

All four signal choices lower mean annual RMSE relative to the shared background for every crop. LAI reductions range from 0.70% to 3.41%, with lower error in at least eleven of thirteen years per crop. NDVI+GPP attains the lowest error for maize and soybean, reducing the background error by 4.69% and 5.32%; NDVI alone is lowest for rice and wheat, with reductions of 3.20% and 4.79%. Additional products therefore need not improve every crop. This ranking concerns observed inputs in a direct model, not the accuracy or downstream value of their predicted unobserved windows.

Figure 7 in the main text summarizes these matched input-removal and product comparisons.

D.5 HISTORICAL-ANCHOR AND READOUT-DISTRIBUTION CONTROLS

Trained-epoch-only historical-anchor control. This diagnostic uses the same temporal blocks, crop recipes, and model families, but requires the historical expert to have at least one trained epoch. Each component is reinitialized with seed 42 and fitted through the full cutoff for the selected duration. The historical MLP has widths 20–256–128–1, GELU activations and dropout 0.1. AdamW uses learning rate $3 \cdot 10^{-4}$, weight decay 10^{-4}, batch size 2048, at most 50 epochs and patience 8. Soybean retains a subsequent TabM residual expert with 16 members, three width-128 blocks, dropout 0.1, learning rate 10^{-3}, batch size 1024, at most 60 epochs and patience 8. Only trained epochs are eligible. The original crop-specific historical branches and maize's equal branch averaging are unchanged. This tests sensitivity to the historical-anchor realization. The refits use Python 3.9.23, PyTorch 2.5.1 with CUDA 12.4, LightGBM 4.6.0 and scikit-learn 1.6.1.

Separating training window from input adaptation. The observed-training readout first uses the original full fitting window, then the common 1993-onward terminal window available for causal state caches. Historical experts and feature statistics use their full preceding windows in both cases. Observed-window-only, historical-mean-completed and BIID-completed readouts use the common window and identical 501/537-dimensional feature layouts. For each 10%, 30% and 50% unobserved window, terminal LightGBM uses 15 leaves, minimum leaf size 200, $L_2 = 10$, learning rate 0.03, at most 1,200 trees, patience 60, year-balanced training weights and mean annual internal-validation RMSE. After preceding-validation selection, the readout is freshly initialized and refitted through the block cutoff with the selected tree count. Historical predictions on terminal training rows remain in-sample; this experiment changes terminal state-input distributions, not historical cross-fitting. It comprises 24 observed-training and 108 adapted logical readouts, with 165 full-fit tree branches because maize averages two branches.

The trained-epoch-only control isolates the influence of the historical-anchor realization. For soybean, the observed-trajectory correction reduces the refitted anchor's mean annual RMSE from 0.3544 to 0.3440. The common-window controls then separate terminal fitting coverage from the state-input

distribution while retaining the same crop recipes. Together, these controls identify historical-anchor quality and terminal fitting coverage as important components of the state-to-yield route.

At a 10% unobserved window, BIID-adapted heads improve over the strong historical reference by 3.53% and 2.50% for maize and rice. Comparisons with the fixed-head completion variants further distinguish the contribution of readout adaptation from that of trajectory completion, providing two concrete levers for configuring the crop-specific terminal predictor.

Causal state reuse. Training trajectories from 1993 onward are generated by weights fitted strictly before their target year. Outer evaluation uses exact state cutoffs 2001, 2005 and 2009. Internal validation uses the latest existing weights no later than the internal fitting cutoff: 1998, 2001 and 2007, respectively. All reused state models have at least one trained epoch. These diagnostics fix state seed 42 and do not measure random-seed uncertainty.

Simple state controls and anomaly fidelity. Tables 19–20 compare identical hidden, crop-active, finite-target slots. Previous-trajectory completion uses the preceding input trajectory with grid–month historical mean for missing values. Persistence holds the last finite visible value; without a visible value it uses the last finite preceding-trajectory value, then historical mean. It never reads the hidden target. For anomaly correlation, both target and forecast subtract the same fitted grid–month historical mean before within-year Pearson correlation; annual correlations are then averaged. Historical mean's constant zero anomaly has undefined correlation, rather than perfect correlation.

BIID has higher anomaly correlation than copying the previous trajectory across the reported conditions, and also lower total state RMSE than historical-mean, previous-year, and last-observation completion. These results show that BIID better preserves both absolute vegetation states and interannual anomalies. The accompanying terminal controls assess how this state information contributes to yield prediction, so the two evaluations provide complementary evidence for the complete pipeline.

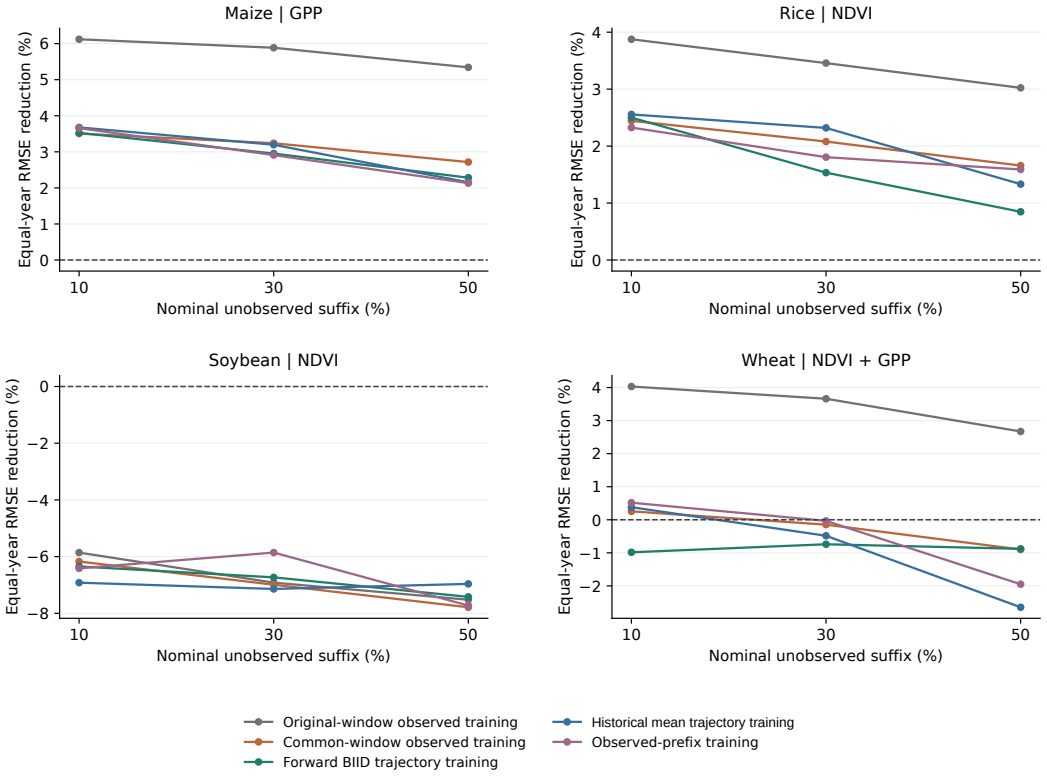

Figure 21: RMSE reductions versus the historical reference under historical-anchor and readout-distribution controls.

D.6 DIRECTIONAL BIID ABLATION

This ablation changes only the interaction inside each state-transition block. *Concat* replaces relevance modulation by parameter-matched bidirectional pooled concatenation. The two directional variants retain only Weather→State or State→Weather relevance modulation; the inactive direction is an identity mapping. Full BIID retains both directions. All variants preserve the retained state tokens, two interaction layers, subsequent cross-attention and gate, optimization budget, temporal blocks, and seed 42.

Table 2(a) tests whether bidirectional interaction improves hidden vegetation dynamics. Each entry first computes physical-unit hidden-slot RMSE within each evaluation year, then averages the thirteen annual scores and the 10%, 30%, and 50% unobserved windows. NDVI averages rice, soybean, and wheat with equal crop–product weight; GPP averages maize and wheat. The two products remain separate because their units differ. Full BIID is the same registered BIID completion as Tables 11, 18, and 19. It reuses those checkpoints and trajectories, so its NDVI and GPP entries are the Table 19 BIID column aggregated with these weights, and its yield entries are the corresponding Table 18 BIID + observed-feedback row averages. Concat and the two one-directional variants newly train only the interaction replacement; they do not reuse those registered trajectories. Panel (b) tests whether any state improvement transfers to yield. Each completed trajectory is passed through the same frozen crop-specific history experts and yield heads. Terminal weather, observed-window inputs, labels, and evaluation samples are unchanged. Entries average the same three windows after computing the thirteen-year mean annual spatial yield RMSE for each crop. Full BIID is lowest on both vegetation products and on all four crop yields.