GEWORLD: A GENOTYPE-ENVIRONMENT WORLD MODEL FOR CROP PHENOTYPE PREDICTION AND ROLLOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Phenotype prediction aims to predict traits from genotype and environmental information. Most existing methods learn static mappings from aggregated data, making it difficult to dynamically adapt to new observations or changing environments. Here, we propose GEWorld, a genotype–environment world model that casts phenotype prediction as the evolution of a latent physiological state: it fuses genotypes with an environmental embedding through a gated, multiplicative genotype-by-environment interaction to initialize this state, assimilates observed phenotype prefixes through a GRU-based observation module to correct the state as new measurements arrive, and evolves the state open-loop through a Mamba-2-inspired state-space transition under future environments and elapsed time to project how the trajectory unfolds. GEWorld decodes static terminal traits from the initial state and residual dynamic trajectories from the rolled states; environmental counterfactuals are obtained by rolling the same assimilated state under alternative future weather. Across static and dynamic benchmarks spanning five crops, GEWorld achieves state-of-the-art predictive accuracy, while its forecasted phenotype trajectories closely track observed development. The code is publicly available at `https://anonymous.4open.science/r/GEWorld-E070/`.

1 INTRODUCTION

Phenotype prediction (PP) aims to infer observable traits from genotype and environmental conditions, with applications ranging from precision medicine to agricultural breeding. In crops, PP leverages the interplay between genotype and agronomic drivers (e.g., soil, moisture, light, and temperature) to predict varietal performance across growing conditions and to accelerate selection (Bailey-Serres et al., 2019; Sun et al., 2026). A central challenge is modeling genotype-by-environment (G×E) interactions, which characterize how the phenotypic effects of a genotype vary across environments (Crossa et al., 2022; Namba et al., 2026).

However, most existing PP methods nevertheless cast the problem as a static mapping from aggregated genotype–environment inputs to phenotypes. Linear mixed models (e.g., GBLUP) (Habier et al., 2007) and reaction-norm models (Jarquín et al., 2014) capture G×E via bilinear or kernelized terms, with BGGE-style kernels extending to multi-environment settings (Granato et al., 2018). Recent deep learning approaches improve modeling flexibility through feature fusion, attention mechanisms, and mixture-of-experts architectures (Zhang et al., 2026; Yao et al., 2025). However, these approaches typically learn input–output associations and do not maintain an explicit latent state that can be updated as new measurements arrive and rolled forward under future, time-varying environments. In contrast, process-based crop models can generate trajectories (Holzworth et al., 2014; Gavasso-Rita et al., 2024), but they rely on strong physiological assumptions and calibration, limiting scalability across large genomic populations. Thus, existing methods still lack an explicit, rollable state that describes how a trait evolves through time.

These gaps motivate a dynamical view of G×E, where genotype modulates the evolution of an underlying physiological state; temporally varying environments drive state transitions; and sparse phenotype measurements provide partial observations of this latent trajectory (Van Voorn et al.,

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.

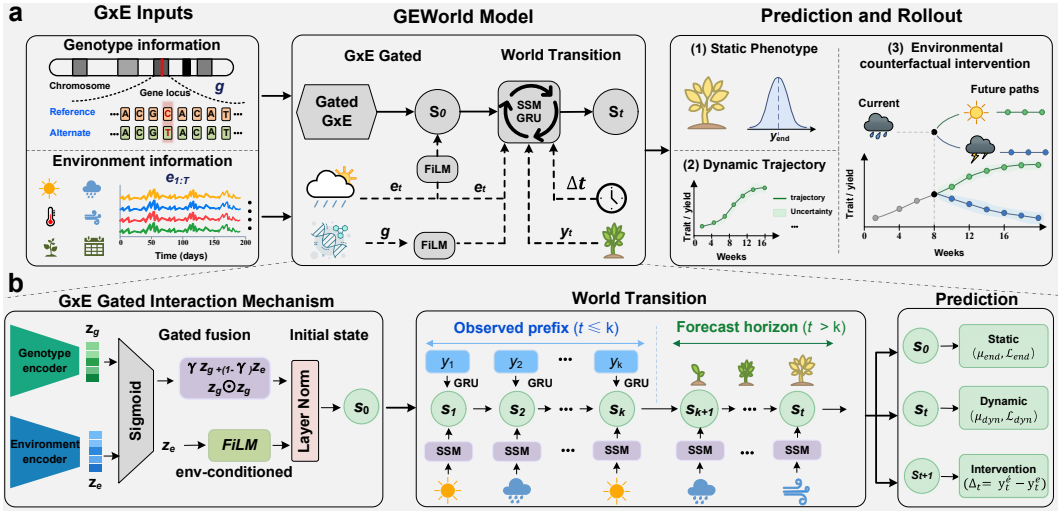


Figure 1: Overview of GEWorld. **(a)** Given genotype g and environmental covariates $e_{1:T}$, a gated multiplicative G×E encoder fuses g with an environmental embedding to initialize the latent physiological state s_0. A GRU observation unit assimilates phenotype prefixes, and a Mamba-2-inspired state-space transition evolves the state under future environments and elapsed time for dynamic forecasting and environment-counterfactual rollouts. **(b)** Module-level details of the G×E initialization, the assimilation–transition loop, and the decoding heads used for phenotype prediction.

2023). This view calls for an evolving latent state that can assimilate incoming observations and propagate through learned dynamics under future environmental inputs.

We introduce GEWorld, a genotype–environment world model for multi-crop phenotype prediction and rollout, as illustrated in Figure 1. GEWorld initializes a latent physiological state via a gated multiplicative G×E fusion of additive SNP-dosage genotypes and environmental embeddings. At each observed time step, a Mamba-2-inspired state-space transition advances the state conditioned on environment and elapsed time, and a GRU-based observation module assimilates the phenotype measurement to correct the prediction. The resulting state supports endpoint prediction of static terminal traits and rollout of residual dynamic trajectories; environment counterfactuals are obtained by rolling the same assimilated state under alternative future weather. The main contributions are summarized as follows:

- **World-model formulation for phenotype prediction.** We cast PP as a partially observed dynamical system with G×E-modulated latent physiology: genotype parameterizes the state evolution, time-varying environments drive the transitions, and phenotype measurements provide sparse observations. We instantiate this view with GEWorld, a rollable genotype–environment world model that supports temporal extrapolation beyond static genotype–environment associations.

- **End-to-end assimilation and rollout learning.** We introduce an end-to-end training framework that couples online state assimilation with environment-conditioned forecasting. Phenotype prefixes are assimilated to correct the latent state, which is then propagated under future environmental covariates and elapsed time. Separate decoding heads produce (i) static phenotype prediction via a Gaussian head on the initialized state and (ii) dynamic trajectory forecasting via an adaptive trajectory decoder on rolled states.

- **Multi-crop evaluation and counterfactual forecasting.** We benchmark GEWorld on five crop species under both static and dynamic prediction. GEWorld achieves state-of-the-art dynamic trajectory forecasting and improves forecasting across multiple observation-prefix ratios on Sorghum and Wheat. Rolling the same assimilated state under alternative future weather yields counterfactual trajectories that recover cross-environment phenotype trends.

2 RELATED WORK

2.1 CLASSICAL G×E MODELING

Genomic prediction aims to predict field phenotypes from genome-wide markers, enabling earlier selection and shorter breeding cycles (Crossa et al., 2017; Wang et al., 2023). Since observed traits arise from genotype–environment interplay, improving robustness across environments has become a core objective in modern breeding (Crossa et al., 2022). Classical statistical models, including GBLUP (Habier et al., 2007) and reaction-norm formulations (Jarquín et al., 2014), represent G×E by combining genomic relationship structure with environment-specific effects. Extensions such as BGGE (Granato et al., 2018) and BMTME (Montesinos-López et al., 2016) further generalize this paradigm to multi-environment and multi-trait settings. While interpretable and empirically strong, these methods typically treat G×E as a static interaction term for endpoint prediction, rather than as a temporally evolving process driven by within-season environmental variation. As a result, they lack an explicit, rollable state representation that could be updated with intermediate measurements and propagated to forecast phenotypic trajectories.

2.2 DEEP LEARNING FOR PHENOTYPE PREDICTION

Deep learning has become a popular alternative to classical G×E models for capturing nonlinear genomic effects and complex cross-modal interactions (Khaki & Wang, 2019). Many recent predictors jointly encode genotypes and environments using feature fusion modules, attention, or mixture-of-experts to improve generalization across trials (Zhang et al., 2026; Yao et al., 2025; Zhou et al., 2026). In parallel, DNA foundation models provide transferable sequence-derived representations that can be integrated into phenotype prediction pipelines (Brixi et al., 2026); for instance, DNABERT-2-based SNP-context and whole-genome embeddings improve prediction in rice and maize relative to marker-only inputs (Li et al., 2026), and PlantCaduceus targets cross-species functional annotation and variant-effect modeling (Zhai et al., 2025). Despite this increased representational capacity, most deep PP methods still operate as static regressors from aggregated genotype–environment descriptors to endpoint traits, with limited support for (i) assimilating intermediate phenotype measurements and (ii) rolling forward under future, time-varying environments to forecast trajectories. In contrast, GEWorld explicitly models phenotype development as latent-state dynamics conditioned on genotype and driven by temporal environmental inputs, enabling online updates and environment-conditioned rollouts.

2.3 CROP GROWTH MODELS AND WORLD MODELS

Process-based crop growth models simulate development via explicit physiological mechanisms and can generate within-season trajectories under environmental drivers such as radiation, temperature, and water availability (Van Voorn et al., 2023; Hajjarpoor et al., 2022). Their mechanistic nature improves interpretability, but it also requires extensive assumptions, parameterization, and calibration, making it challenging to integrate high-dimensional genomic variation at scale. World models offer a complementary, data-driven alternative: they learn a compact latent state together with observation and transition dynamics directly from data, enabling rollouts under changing exogenous inputs (Hafner et al., 2025; Wei et al., 2026). This paradigm is well suited to partially observed dynamical systems, where an unobserved state evolves, and measurements are sparse and noisy. Crop phenotypes naturally fit this template, where genotype modulates the dynamics, temporally varying environments drive state transitions, and field measurements provide partial observations of the evolving trajectory. These considerations motivate genotype-conditioned world modeling for dynamic G×E phenotype prediction and counterfactual simulation.

3 THE PROPOSED GEWORLD

GEWorld casts phenotype prediction as the evolution of a latent physiological state: the state is initialized from genotype and trial-level context, corrected when measurements become available, and advanced under time-resolved environmental inputs. Static phenotype prediction, dynamic trajectory rollout, and environment counterfactual simulation then emerge as different readouts of the same underlying state process (Figure 1).

3.1 PROBLEM FORMULATION

We study phenotype prediction from genotype and environment under both static and dynamical observation protocols. Each sample is represented as a genotype–environment trajectory

$$\mathcal{T} = \left(\mathbf{g}, \mathbf{e}_{1:T}, \mathbf{y}^{\mathrm{end}}, \mathbf{y}^{\mathrm{dyn}}_{1:T}, \mathbf{m}_{1:T}\right), \tag{1}$$

where $\mathbf{g} \in \mathbb{R}^M$ is the additive dosage vector over M genetic markers and $\mathbf{e}_{1:T}$ is the sequence of time-resolved environmental covariates over a season of length T. On the phenotype side, $\mathbf{y}^{\mathrm{end}}$ denotes the terminal (endpoint) trait, $\mathbf{y}^{\mathrm{dyn}}_{1:T}$ denotes the within-season phenotype trajectory, and $\mathbf{m}_t \in \{0,1\}^C$ is a per-trait observation mask indicating which of the C traits are observed at time step t.

We consider three prediction settings that differ in the phenotype information available at prediction time and in how the future environment is specified. **(i) Static phenotype prediction** estimates $\mathbf{y}^{\mathrm{end}}$ from genotype $\mathbf{g}$ and a trial-level environmental summary $\mathbf{e}$ computed from $\mathbf{e}_{1:T}$, without observing within-season phenotypes. **(ii) Dynamic trajectory rollout** additionally conditions on an observed phenotype prefix $\mathbf{y}^{\mathrm{dyn}}_{1:k}$ (with mask $\mathbf{m}_{1:k}$) and the realized environment up to the cutoff k, and predicts the remaining trajectory $\mathbf{y}^{\mathrm{dyn}}_{k+1:T}$. **(iii) Counterfactual forecasting** retains the genotype and observed phenotype prefix, then branches from the same assimilated state at cutoff k. One branch is rolled out under the factual future environment $\mathbf{e}_{k+1:T}$, while the other uses an alternative future sequence $\mathbf{e}'_{k+1:T}$. The difference between their predicted phenotype trajectories is considered as the model-predicted response to the environmental substitution.

3.2 GATED G×E INITIALIZATION

The initial state $\mathbf{s}_0$ is designed to compactly summarize genotype, environment, and their interaction in a shared latent space. We first embed the additive marker dosage vector $\mathbf{g}$ and a trial-level environmental summary $\mathbf{e}$ via separate MLP encoders,

$$\mathbf{z}_g = \mathrm{MLP}_g(\mathbf{g}), \qquad \mathbf{z}_e = \mathrm{MLP}_e(\mathbf{e}), \qquad \mathbf{z}_g, \mathbf{z}_e \in \mathbb{R}^d. \tag{2}$$

We then fuse the two embeddings with a dimension-wise gate that adaptively balances genetic and environmental contributions,

$$\boldsymbol{\gamma} = \sigma(\mathbf{W}[\mathbf{z}_g; \mathbf{z}_e] + \mathbf{b}), \qquad \mathbf{u} = \mathrm{LayerNorm}(\boldsymbol{\gamma} \odot \mathbf{z}_g + (\mathbf{1} - \boldsymbol{\gamma}) \odot \mathbf{z}_e). \tag{3}$$

Here, $[\cdot\,;\cdot]$ denotes concatenation and $\odot$ denotes elementwise multiplication. To explicitly encode G×E effects, we concatenate the gated fusion $\mathbf{u}$ with the multiplicative interaction $\mathbf{z}_g \odot \mathbf{z}_e$, map the result to an intermediate state $\widetilde{\mathbf{s}}$, and apply an environment-conditioned FiLM modulation (Perez et al., 2018),

$$\begin{aligned} \widetilde{\mathbf{s}} &= \mathrm{MLP}_s([\mathbf{u}; \mathbf{z}_g \odot \mathbf{z}_e]), \ (\boldsymbol{\phi}_e, \boldsymbol{\psi}_e) = \mathrm{Linear}(\mathbf{z}_e), \\ \mathbf{s}_0 &= \mathrm{LayerNorm}(\widetilde{\mathbf{s}} \odot [\mathbf{1} + \alpha \tanh(\boldsymbol{\phi}_e)] + \alpha \boldsymbol{\psi}_e). \end{aligned} \tag{4}$$

Scale and shift $(\boldsymbol{\phi}_e, \boldsymbol{\psi}_e)$ are derived from $\mathbf{z}_e$; the $\tanh$-bounded scaling and modulation strength α control the magnitude of environment-dependent adaptation at initialization.

3.3 OBSERVATION ASSIMILATION AND STATE TRANSITION

Starting from the initialized state $\mathbf{s}_0$, GEWorld advances the latent state one interval at a time under the time-resolved environment and is corrected by phenotype measurements on the observed prefix. For $t \geq 1$, we first compute a transition-predicted state $\bar{\mathbf{s}}_t$ using a one-step selective state-space transition,

$$\bar{\mathbf{s}}_t = \mathrm{FiLM}_g\Big(\mathrm{LayerNorm}\Big(\mathrm{SSM}\big([\mathbf{s}_{t-1};\, \mathrm{MLP}_e(\mathbf{e}_t);\, \tau(\Delta t_t)], \mathbf{s}_{t-1}\big)\Big)\Big), \tag{5}$$

where MLP_e encodes the standardized environmental covariates at step t and τ is a two-layer MLP that embeds the elapsed interval Δt_t. The FiLM operator FiLM_g modulates the transition, enabling genotype-specific dynamics.

The transition core SSM is a recurrent, Mamba-2-inspired selective state-space cell derived from a continuous controlled linear system. Let $\mathbf{x}_t$ denote its input. SSM maintains a projected state memory $\mathbf{q}_t$ updated as

$$\mathbf{q}_t = \exp(\boldsymbol{\Delta}_t \odot \mathbf{A}) \odot P(\mathbf{s}_{t-1}) + \boldsymbol{\Delta}_t \odot \mathbf{B}_t \odot \mathbf{u}_t, \qquad \mathbf{A} = -\exp(\mathbf{A}_{\log}), \tag{6}$$

where $\mathbf{u}_t$ is the transformed transition input (derived from $\mathbf{x}_t$) and $\mathbf{A}$ is a learned stable state matrix. The memory is read out through $\mathbf{C}_t$ and projected back to the latent dimension to produce the transition output. We implement SSM as a one-step recurrence to support open-loop rollouts.

At each time step with a valid phenotype measurement, we correct the transition-predicted state using a GRU update.

$$\mathbf{s}_t^{\text{corr}} = \text{LayerNorm}\Big(\text{GRUCell}\big(\text{MLP}_o(\mathbf{y}_t^{\text{dyn}}; \mathbf{e}_t), \bar{\mathbf{s}}_t\big)\Big), \qquad \mathbf{s}_t = \begin{cases} \mathbf{s}_t^{\text{corr}}, & m_t = 1, \\ \bar{\mathbf{s}}_t, & m_t = 0. \end{cases} \tag{7}$$

where $m_t \in \{0, 1\}$ indicates whether a valid phenotype observation is present at step t. Here, $\bar{\mathbf{s}}_t$ is the transition-predicted state and $\mathbf{s}_t$ is the observation-corrected state. The GRU is a gated recurrent assimilation mechanism (Cho et al., 2014). Assimilation is applied only on observed steps up to the cutoff k; for $t > k$, the model runs open-loop and propagates the state purely via the transition under the specified forecast environmental sequence.

3.4 Training objectives

(i) Static phenotype prediction: The terminal phenotype is decoded directly from the initial state using a heteroscedastic Gaussian head,

$$(\mu^{\text{end}}, \ell^{\text{end}}) = \text{MLP}_{\mu\ell}(\mathbf{s}_0), \qquad \mathcal{L}_{\text{end}} = \frac{1}{2}\left[\ell^{\text{end}} + \big(\mathbf{y}^{\text{end}} - \mu^{\text{end}}\big)^2 \exp\big(-\ell^{\text{end}}\big)\right], \tag{8}$$

Here, μ^{end} is the vector of predicted endpoint means, and $\ell^{\text{end}} = \log \sigma^2$ is the predicted log-variance, enabling input-dependent uncertainty.

(ii) Dynamic trajectory rollout: We use an adaptive decoder to predict the phenotype trajectory from the latent states $\{\mathbf{s}_t\}$. Its residual branch predicts a residual change from the preceding phenotype:

$$\mu_t^{\text{res}} = \hat{\mathbf{y}}_{t-1} + \text{MLP}_{\text{dyn}}(\mathbf{s}_t). \tag{9}$$

Here, μ_t^{res} is the phenotype prediction from the residual branch before adaptive mixing and $\hat{\mathbf{y}}_{t-1}$ denotes the last available phenotype value.

To limit drift during long-horizon rollouts, a reference branch predicts a bounded state-dependent correction to a base trajectory:

$$\mathbf{b}_t = (1 - \eta)\hat{\mathbf{y}}_{t-1} + \eta \mathbf{c}_t, \qquad \mu_t^{\text{ref}} = \mathbf{b}_t + \tanh(\text{MLP}_\Delta(\mathbf{s}_t)), \tag{10}$$

where $\mathbf{c}_t$ is the training-set mean phenotype at stage t, $\hat{\mathbf{y}}_{t-1}$ is the preceding observed or predicted phenotype, and η is a learned scalar mixing weight. The final prediction combines this reference branch with the residual branch using a learned per-trait gate:

$$\boldsymbol{\rho} = \sigma(\boldsymbol{\ell}_\rho), \qquad \mu_t^{\text{dyn}} = (\mathbf{1} - \boldsymbol{\rho}) \odot \mu_t^{\text{res}} + \boldsymbol{\rho} \odot \mu_t^{\text{ref}}, \tag{11}$$

where $\boldsymbol{\ell}_\rho$ is initialized so that the reference branch has a small initial weight. The gate is shared across forecast steps. The trajectory loss is evaluated over observed step–trait pairs after the cutoff,

$$\mathcal{L}_{\text{dyn}} = \frac{\sum_{(t,j)\in\mathcal{V}} w_j \left(\mathbf{y}_{t,j}^{\text{dyn}} - \mu_{t,j}^{\text{dyn}}\right)^2}{\sum_{(t,j)\in\mathcal{V}} w_j}, \tag{12}$$

where $\mathcal{V} = \{(t, j) : t > k,\ m_{t,j} = 1\}$ indexes valid observations and w_j balances traits with different numerical scales.

(iii) Counterfactual inference: Given an observed prefix up to cutoff k, we first run GEWorld in closed loop under the factual environment to assimilate observations and obtain the posterior state $\mathbf{s}_k$. We then perform two open-loop rollouts from the same state $\mathbf{s}_k$: a factual rollout driven by the observed future environment $\mathbf{e}_{k+1:T}$ and a counterfactual rollout driven by the alternative sequence $\mathbf{e}'_{k+1:T}$. Denoting the resulting predictions by

$$\hat{\mathbf{y}}_t^{\text{f}} = \hat{\mathbf{y}}_t(\mathbf{e}_{k+1:T}), \qquad \hat{\mathbf{y}}_t^{\text{cf}} = \hat{\mathbf{y}}_t(\mathbf{e}'_{k+1:T}), \qquad \hat{\boldsymbol{\Delta}}_t = \hat{\mathbf{y}}_t^{\text{cf}} - \hat{\mathbf{y}}_t^{\text{f}}, \qquad t > k, \tag{13}$$

where $\hat{\mathbf{y}}_t^{\text{f}}$ and $\hat{\mathbf{y}}_t^{\text{cf}}$ denote the predicted phenotype vectors at time t under the factual and counterfactual future environments, respectively. Thus, $\hat{\boldsymbol{\Delta}}_t$ measures the model-implied response to replacing the future environment while holding genotype and the assimilated state fixed.

4 EXPERIMENTS

We evaluate GEWorld with four questions.

- **RQ1**: How well does it predict phenotype versus classical and deep baselines?
- **RQ2**: How well does open-loop prediction work after prefix assimilation?
- **RQ3**: Do counterfactual rollouts with swapped future weather match observed phenotype shifts?
- **RQ4**: What do ablation, embedding, and robustness analyses reveal about the model?

4.1 EXPERIMENTAL SETTINGS

Benchmark Datasets and Model: We evaluate GEWorld under two complementary settings: *static* phenotype prediction and *dynamic* trajectory rollout (Table S1). The static benchmark comprises five datasets; each sample contains an SNP-dosage vector and trial-level environmental covariates. The dynamic benchmark contains time series for sorghum and wheat, including weekly APSIM simulations and real sensor measurements from TERRA-REF and the Field Phenotyping Platform, indexed by genotype, environment, and time. We compare against a diverse set of baselines. Static baselines include GBLUP (Habier et al., 2007), BGGE (Granato et al., 2018), MoEGE2P (Zhang et al., 2026), KIBREED (Gogna et al., 2026), GEFormer (Yao et al., 2025), and GE-BiFormer (Zhou et al., 2026). Dynamic baselines include Persistence (Hyndman et al., 2014), HistoryMeanEnv, dynamicGP (Hobby et al., 2025), TCN (Lea et al., 2017), and Autoformer (Wu et al., 2021).

Benchmark evaluation and details: We use genotype-grouped splits shared across methods, with a $70/15/15$ split by genotype ID for train/validation/test. We select hyperparameters on the validation set and report test performance once. For static prediction, models predict endpoint traits from genotype and trial-level environment. For dynamic prediction on sorghum and wheat, models observe a phenotype prefix of $25\%/50\%/75\%$ of the season and forecast the remaining trajectory under known future environments. We report test R^2, PCC, and RMSE for static prediction, and report R^2 over the post-cutoff interval for dynamic forecasting.

4.2 BENCHMARK RESULTS

Static phenotype prediction (RQ1): Table 1 summarizes phenotype prediction accuracy across five crops. GEWorld achieves the best mean performance for every species and metric, with a macro-average R^2 of 0.703, indicating consistently stronger generalization than both classical and deep baselines. The largest performance improvements are observed for Soybean and Rice, where the strongest baselines leave substantial variance unexplained. These results suggest that the advantage arises from the more suitable inductive bias provided by gated multiplicative fusion for integrating the two modalities and adjusting their relative contributions. By contrast, gains are smaller on Maize

Table 1: Static phenotype prediction results. Higher R^2 and PCC, and lower RMSE, indicate better performance. Best results are in bold and second-best results are underlined.

Species	Metric	GBLUP	BGGE	MoEGE2P	KIBREED	GEFormer	GE-BiFormer	**GEWorld**
Maize	R^2	0.603±0.008	0.616±0.006	0.608±0.005	0.616±0.006	0.634±0.008	0.613±0.010	**0.661±0.005**
	PCC	0.777±0.005	0.785±0.004	0.782±0.004	0.786±0.004	0.798±0.005	0.783±0.006	**0.813±0.003**
	RMSE	1.947±0.030	1.916±0.021	1.934±0.022	1.915±0.026	1.870±0.029	1.923±0.034	**1.800±0.020**
Soybean	R^2	0.666±0.013	0.682±0.013	0.572±0.037	0.571±0.022	0.323±0.013	0.588±0.130	**0.761±0.008**
	PCC	0.816±0.008	0.826±0.008	0.761±0.022	0.758±0.014	0.572±0.012	0.764±0.093	**0.873±0.004**
	RMSE	645.17±7.72	629.65±10.42	729.34±21.01	731.10±23.41	918.03±12.30	710.90±114.98	**545.28±9.27**
Sorghum	R^2	0.517±0.057	0.537±0.077	0.532±0.079	0.462±0.062	0.494±0.115	0.468±0.107	**0.560±0.080**
	PCC	0.725±0.042	0.741±0.054	0.740±0.048	0.691±0.048	0.717±0.081	0.693±0.072	**0.755±0.051**
	RMSE	5634.1±884.1	5509.3±938.5	5530.7±901.7	5939.3±886.2	5757.6±1168.1	5889.1±1018.0	**5372.7±964.8**
Rice	R^2	0.460±0.071	0.467±0.076	0.484±0.069	0.505±0.079	0.510±0.041	0.441±0.103	**0.574±0.068**
	PCC	0.690±0.040	0.696±0.042	0.703±0.042	0.724±0.046	0.719±0.032	0.687±0.064	**0.764±0.043**
	RMSE	10.426±0.852	10.357±0.817	10.181±0.686	9.990±0.993	9.966±0.935	10.618±1.380	**9.244±0.881**
Wheat	R^2	0.951±0.002	0.954±0.002	0.940±0.003	0.941±0.004	0.953±0.003	0.949±0.003	**0.958±0.002**
	PCC	0.975±0.001	0.977±0.001	0.970±0.001	0.971±0.002	0.976±0.002	0.974±0.001	**0.979±0.001**
	RMSE	0.310±0.006	0.303±0.007	0.344±0.006	0.340±0.012	0.305±0.008	0.319±0.008	**0.288±0.007**

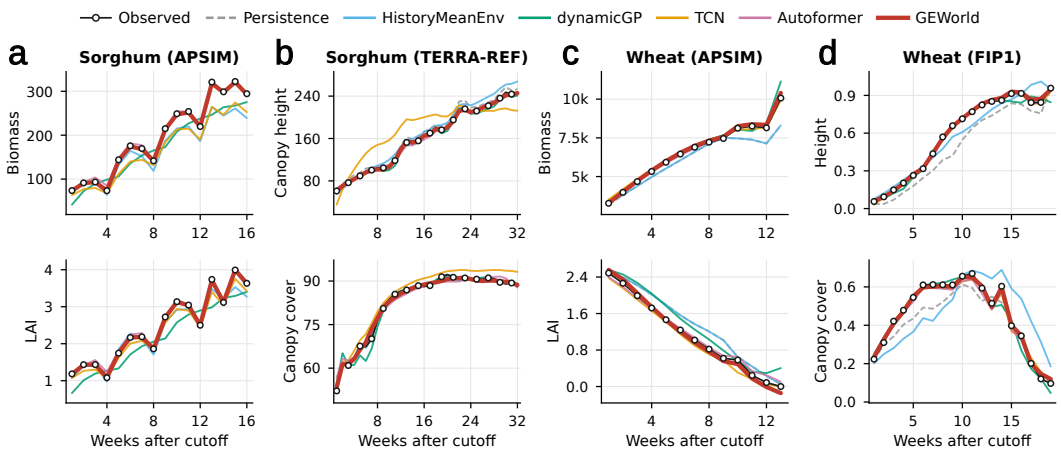


Figure 2: **(a)–(d)** Dynamic trajectory forecasting with a 50% observed-prefix cutoff. Columns correspond to four datasets: simulated sorghum, real sorghum, simulated wheat, and real wheat. Curves report the mean prediction over test trajectories and random seeds.

and on Wheat, where performance is near saturation. This pattern is consistent with diminishing returns when the signal-to-noise ratio is high or when the benchmark approaches a performance ceiling; additional modeling capacity yields limited headroom.

Dynamical trajectory rollouts (RQ2): We evaluate open-loop trajectory forecasting on sorghum and wheat under *known* future environments, varying the amount of observed within-season data (25%/50%/75% prefix). All methods receive the same realized future weather, and performance differences reflect how well models infer latent developmental stage from the prefix and propagate it forward. On real field trajectories (Table 2), GEWorld performs strongest in challenging long-horizon regimes, including Sorghum cover at the 75% cutoff and Wheat height at the 75% cutoff. Because late-season dynamics are often nonlinear and trait-specific, strong performance suggests that GEWorld does more than persistence-style extrapolation and can maintain stable rollouts when the remaining horizon is short, but the trajectory is difficult to extend. On simulated APSIM trajectories, most traits approach saturation across methods, leaving limited headroom. A more discriminative regime is simulated Wheat LAI with only a 25% prefix (Table S4), where GEWorld achieves $R^2 = 0.916$ versus 0.832 for Autoformer and ≈ 0.23 for Persistence and TCN. This large gap indicates that the learned latent state is particularly valuable when observations are sparse, and the developmental phase is ambiguous. Assimilation can anchor the hidden physiological state early, after which the genotype- and environment-conditioned transition supports accurate long-horizon rollouts. Figure 2 visualizes representative 50%-prefix rollouts, with complete results for all cutoffs provided in Appendix Figures S1 and S2.

Environment counterfactual intervention (RQ3): We ask whether GEWorld captures the *direction* of cross-environment phenotype changes under explicit environmental interventions. On sorghum and wheat, we assimilate a 50% within-season phenotype prefix under the factual environment e to obtain an identical posterior state $\mathbf{s}_k$ for both rollouts. We then propagate $\mathbf{s}_k$ open-loop under the factual future weather $e_{k+1:T}$ and the counterfactual future weather $e'_{k+1:T}$ using the same learned transition and decoding heads. We summarize the intervention effect by the predicted terminal difference $\hat{\Delta}$ and compare it to the observed endpoint difference Δ between trials (Table S8). Across environment swaps, GEWorld correctly identifies whether the substitution is beneficial or detrimental and provides high rank consistency between predicted and observed effects. Because genotype, prefix observations, and model parameters are held fixed across the two rollouts, these changes can be attributed to the substituted future environment. Figure 3 further visualizes the mechanism: counterfactual trajectories gradually shift toward the trajectories observed in the target environment, indicating that future weather drives coherent changes in the learned latent-state evolution rather than merely perturbing the final readout.

Table 2: Real dynamic trajectory forecasting under known future environmental conditions. Test R^2 is reported for post-cutoff predictions at 25%, 50%, and 75% observed-prefix lengths. Bold and underlined values indicate the best and second-best results in each row, respectively.

Crop	Trait	Prefix	Persistence	HistoryMeanEnv	dynamicGP	TCN	Autoformer	GEWorld
Sorghum (Real)	Height	25%	0.900±0.006	0.875±0.028	0.932±0.007	0.901±0.006	0.908±0.006	**0.951±0.003**
		50%	0.892±0.009	0.858±0.031	0.870±0.013	0.633±0.080	0.893±0.009	**0.922±0.008**
		75%	0.658±0.019	0.663±0.058	0.657±0.028	0.557±0.031	0.633±0.049	**0.671±0.023**
	Cover	25%	0.901±0.006	0.888±0.007	0.907±0.008	0.344±0.112	0.922±0.010	**0.948±0.005**
		50%	0.520±0.025	0.520±0.025	0.622±0.045	0.633±0.074	0.701±0.026	**0.738±0.042**
		75%	0.185±0.116	0.022±0.010	−0.465±0.158	0.176±0.128	−0.698±0.448	**0.528±0.100**
Wheat (Real)	Height	25%	0.701±0.001	0.798±0.002	0.918±0.002	0.959±0.003	0.898±0.026	**0.974±0.007**
		50%	0.568±0.003	0.816±0.001	0.896±0.002	0.953±0.003	0.929±0.002	**0.974±0.003**
		75%	−0.178±0.007	0.232±0.011	0.422±0.012	0.823±0.011	0.430±0.017	**0.884±0.017**
	Cover	25%	0.569±0.001	0.387±0.011	0.636±0.004	0.881±0.006	0.766±0.053	**0.899±0.005**
		50%	0.265±0.001	0.175±0.012	0.526±0.004	0.826±0.014	0.861±0.015	**0.865±0.003**
		75%	0.314±0.004	0.593±0.004	0.462±0.007	0.866±0.010	0.892±0.007	**0.910±0.005**

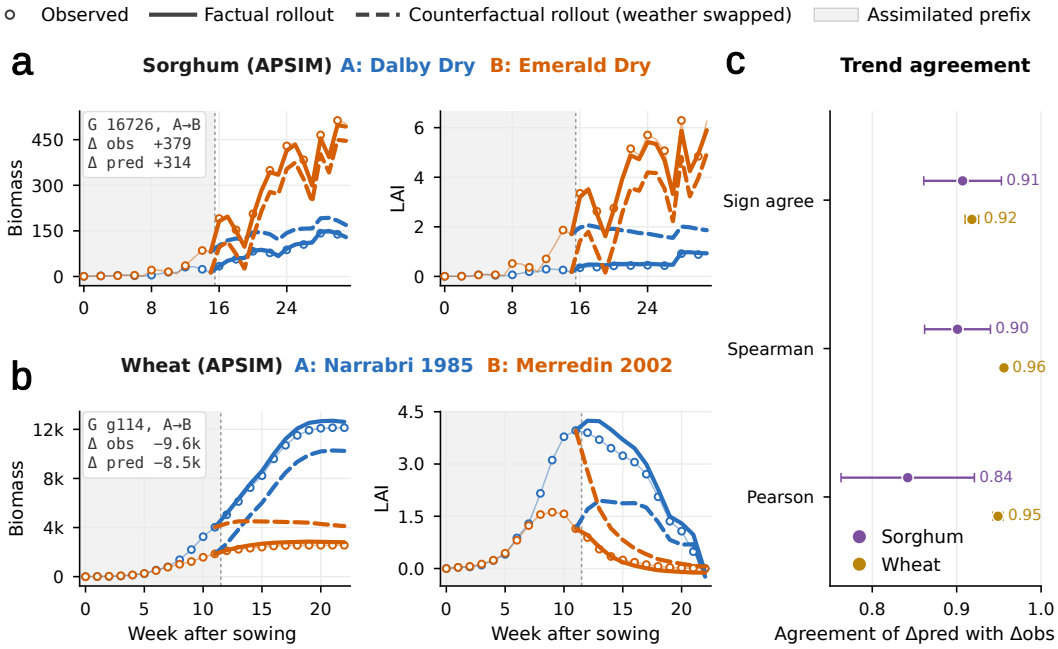


Figure 3: Environmental counterfactual interventions via future-weather substitution. **(a)** Sorghum: factual vs. counterfactual rollouts for biomass and leaf area index under swapped future environments. **(b)** Wheat: corresponding rollouts for biomass and leaf area index. **(c)** Agreement between predicted and observed cross-environment effects, comparing Δ_{pred} to Δ_{obs} using sign consistency as well as Spearman and Pearson correlations.

Ablation analyses (RQ4): We perform two ablation suites. *Static input ablations* on the five-crop benchmark replace gated G×E fusion with simple concatenation (**W/o G×E**) or remove one modality (**W/o genotype**, **W/o environment**). *Dynamic module ablations* on real sorghum (height/cover; 50% and 75% prefixes) disable phenotype assimilation (**W/o Assimilation**), remove the open-loop transition (**W/o Transition**), or replace the adaptive decoder with a residual-only head (**W/o Decode**). For static prediction, Figure 4a shows that removing environment features causes the largest performance drop overall, underscoring that trial-level conditions are indispensable for G×E prediction. Removing genotype is especially damaging for Rice, indicating crop-dependent reliance on genetic signal. Replacing gated fusion with concatenation produces only a moderate decline, indicating that both inputs carry the main predictive signal while the proposed interaction improves their integration. For dynamic forecasting (Figure 4b), the transition module is the dominant fac-

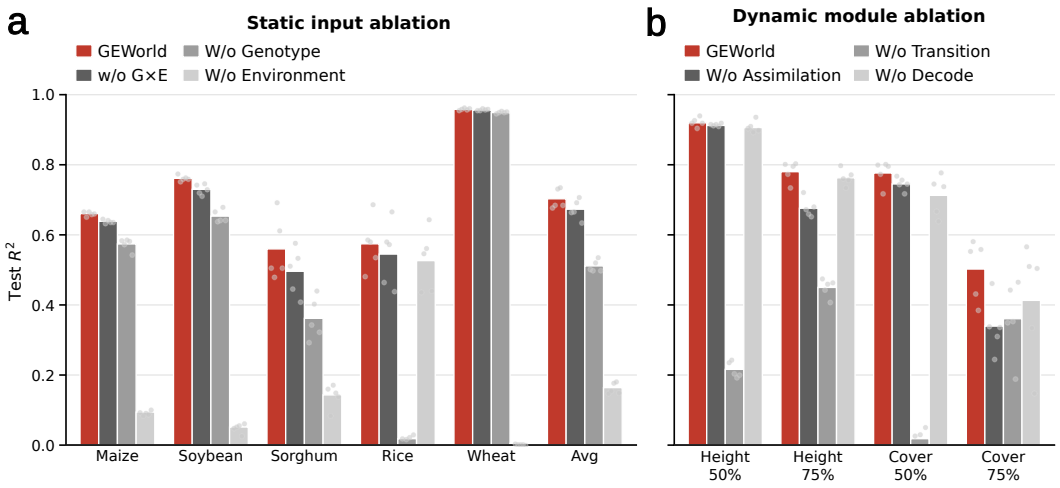


Figure 4: Ablations of GEWorld components. **(a)** Static phenotype prediction on five crops (maize, soybean, sorghum, rice, and wheat). **(b)** Dynamic forecasting on real sorghum, evaluating height and cover rollouts with 50% and 75% observed-prefix cutoffs.

tor: removing open-loop transition collapses 50%-prefix height/cover performance, confirming that accurate rollouts require learned state dynamics rather than direct extrapolation. Assimilation is most beneficial at longer horizons, where correcting the latent state with sparse observations reduces accumulated drift. The adaptive decoder primarily stabilizes difficult traits, complementing the transition by regularizing readout when the forecast horizon grows.

Embedding feature analyses (RQ4): Embedding diagnostics show that the learned gate adapts to the information structure of each crop: sorghum is more environment-leaning, rice is more genotype-leaning, and the remaining crops are closer to balanced. The agreement between gate statistics and occlusion effects suggests that this allocation reflects true predictive dependence rather than incidental parameterization. UMAP projections further indicate that the initialized state $\mathbf{s}_0$ retains phenotype-relevant organization, implying that gated fusion produces a structured joint representation rather than an uninformative mixture. See Appendix C.4 and Figure S4 for details.

Robustness analyses (RQ4): Robustness experiments indicate that performance is generally stable across split protocols but sensitive to environmental support. Accuracy degrades monotonically under joint input corruption, and rare environments or held-out seasons are consistently harder than well-represented conditions, suggesting that limited environmental coverage is the primary practical failure mode. GEWorld maintains the strongest macro performance under phenotype-random splits, although genotype holdout remains the most relevant protocol for breeding deployment. Finally, uncertainty calibration does not always track point accuracy, motivating coverage-aware evaluation and explicit calibration before applying the model to novel environments (Appendix C.5, Figure S5).

5 Conclusion

We presented GEWorld, a genotype–environment world model that treats phenotype prediction as latent-state dynamics: gated G×E fusion initializes a physiological state, GRU assimilation updates it with phenotype prefixes, and a selective state-space transition propagates it under future environments. Endpoint prediction, open-loop trajectory forecasting, and weather-swap counterfactuals are unified as readouts of this shared state process. Across five crops in both simulated and field settings, GEWorld improves genotype-holdout accuracy, strengthens post-prefix forecasting, and captures observed cross-environment trends; ablations further validate the contributions of transition, assimilation, and adaptive decoding. Current limitations include reliance on known future weather and uneven environmental coverage/calibration; future work will integrate environment forecasting and denser longitudinal sensing for decision-oriented breeding.

AI USE STATEMENT

In this work, we used generative AI tools to revise manuscript expressions for clarity and to create or optimize scientific figures and tables. All numerical outputs, analytical results, and scientific claims were independently verified by the authors, who accept full responsibility for the accuracy and integrity of the final work.

ETHICS STATEMENT

This study exclusively analyzes publicly available crop genotype, phenotype, and environmental datasets. It involves no human or animal subjects or experimentation.

REPRODUCIBILITY STATEMENT

All code, preprocessing scripts, and configuration files are publicly available at `https://anonymous.4open.science/r/GEWorld-E070/`. Experiments were conducted using Python 3.10, PyTorch 2.6.0, and standard scientific libraries (NumPy and pandas). Dataset provenance, evaluation protocols, architectural details, and training hyperparameters are provided in Appendices A and B.

REFERENCES

Julia Bailey-Serres, Jane E Parker, Elizabeth A Ainsworth, Giles ED Oldroyd, and Julian I Schroeder. Genetic strategies for improving crop yields. *Nature*, 575(7781):109–118, 2019.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Maxwell Burnette, Rob Kooper, JD Maloney, Gareth S Rohde, Jeffrey A Terstriep, Craig Willis, Noah Fahlgren, Todd Mockler, Maria Newcomb, Vasit Sagan, et al. Terra-ref data processing infrastructure. In *Proceedings of the Practice and Experience on Advanced Research Computing: Seamless Creativity*, pp. 1–7. 2018.

Daniela Bustos-Korts, Marcos Malosetti, Karine Chenu, Scott Chapman, Martin P Boer, Bangyou Zheng, and Fred A Van Eeuwijk. From qtls to adaptation landscapes: using genotype-to-phenotype models to characterize g$\times$ e over time. *Frontiers in Plant Science*, 10:1540, 2019.

Qiuyue Chen, Jacob D Washburn, Dayane Cristina Lima, Maria Cinta Romay, Joseph L Gage, James B Holland, Alencar Xavier, Seth C Murray, David Ertl, Marco Lopez-Cruz, et al. Genomes to fields 2024 maize genotype by environment prediction competition. *BMC Research Notes*, 19 (1):113, 2026.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

José Crossa, Paulino Pérez-Rodríguez, Jaime Cuevas, Osval Montesinos-López, Diego Jarquín, Gustavo De Los Campos, Juan Burgueño, Juan M González-Camacho, Sergio Pérez-Elizalde, Yoseph Beyene, et al. Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, 22(11):961–975, 2017.

José Crossa, Osval Antonio Montesinos-López, Paulino Pérez-Rodríguez, Germano Costa-Neto, Roberto Fritsche-Neto, Rodomiro Ortiz, Johannes WR Martini, Morten Lillemo, Abelardo Montesinos-López, Diego Jarquin, et al. Genome and environment based prediction models and methods of complex traits incorporating genotype$\times$ environment interaction. *Genomic prediction of complex traits: Methods and protocols*, pp. 245–283, 2022.

Yohanne Larissa Gavasso-Rita, Simon Michael Papalexiou, Yanping Li, Amin Elshorbagy, Zhenhua Li, and Corinne Schuster-Wallace. Crop models and their use in assessing crop production and food security: A review. *Food and Energy Security*, 13(1):e503, 2024.

Abhishek Gogna, Bahareh Kamali, Valentin Wimmer, Renate H Schmidt, Ehsan Eyshi Rezaei, Wera Maria Eckhoff, Jochen C Reif, and Yusheng Zhao. Predicting enviromically adapted varieties with big data. *Genome Biology*, 27(1):3, 2026.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Italo Granato, Jaime Cuevas, Francisco Luna-Vázquez, Jose Crossa, Osval Montesinos-López, Juan Burgueño, and Roberto Fritsche-Neto. Bgge: a new package for genomic-enabled prediction incorporating genotype× environment interaction models. *G3: Genes, Genomes, Genetics*, 8(9): 3039–3047, 2018.

David Habier, Rohan L Fernando, and JCM2219482 Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Amir Hajjarpoor, William CD Nelson, and Vincent Vadez. How process-based modeling can help plant breeding deal with g x e x m interactions. *Field Crops Research*, 283:108554, 2022.

David Hobby, Hao Tong, Marc Heuermann, Alain J Mbebi, Roosa AE Laitinen, Matteo Dell'Acqua, Thomas Altmann, and Zoran Nikoloski. Predicting plant trait dynamics from genetic markers. *Nature Plants*, 11(5):1018–1027, 2025.

Dean P Holzworth, Neil I Huth, Peter G deVoil, Eric J Zurcher, Neville I Herrmann, Greg McLean, Karine Chenu, Erik J van Oosterom, Val Snow, Chris Murphy, et al. Apsim–evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62: 327–350, 2014.

Rob J Hyndman, George Athanasopoulos, et al. *Forecasting: principles and practice*, volume 2. OTexts Melbourne, Australia, 2014.

Diego Jarquín, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3):595–607, 2014.

Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10:621, 2019.

Mary-Francis LaPorte, Haixiao Hu, Walter Solomon, James H Oard, Steven J Knapp, Daniel E Runcie, Jeremy D Edwards, Anna McClung, and Christine H Diepenbrock. Multi-environment evaluation and genomic prediction of agronomic traits in the southern us rice genepool. *The Plant Genome*, 19(2):e70222, 2026.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012, 2017.

Huan Li, Yunpeng Cui, Tan Sun, Ting Wang, Zhen Chen, Chao Wang, Wenbo Bian, Juan Liu, Mo Wang, Li Chen, et al. Crop phenotype prediction using snp context and whole-genome feature embedding based on dnabert-2. *Plant Methods*, 22(1):15, 2026.

Osval A Montesinos-López, Abelardo Montesinos-López, José Crossa, Fernando H Toledo, Oscar Pérez-Hernández, Kent M Eskridge, and Jessica Rutkoski. A genomic bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, 6(9):2725–2744, 2016.

Shinichi Namba, Kyuto Sonehara, Yuriko N Koyanagi, Takezo Kikuchi, Takafumi Ojima, Ryuya Edahiro, Go Sato, Taiki Yamaji, Yoshihiko Tomofuji, Hiroyuki Ueda, et al. A cross-population compendium of gene–environment interactions. *Nature*, 651(8106):688–697, 2026.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Paulino Pérez-Rodríguez, José Crossa, Jessica Rutkoski, Jesse Poland, Ravi Singh, Andrés Legarra, Enrique Autrique, Gustavo de los Campos, Juan Burgueño, and Susanne Dreisigacker. Single-step genomic and pedigree genotype× environment interaction models for predicting wheat lines in international environments. *The Plant Genome*, 10(2):plantgenome2016–09, 2017.

Lukas Roth, Mike Boss, Norbert Kirchgessner, Helge Aasen, Brenda Patricia Aguirre-Cuellar, Price Pius Atuah Akiina, Jonas Anderegg, Joaquin Gajardo Castillo, Xiaoran Chen, Simon Corrado, et al. The fip 1.0 data set: Highly resolved annotated image time series of 4,000 wheat plots grown in 6 years. *GigaScience*, 14:giaf051, 2025.

Vitor Seiti Sagae, Moysés Nascimento, Ana Carolina Campana Nascimento, Felipe Lopes da Silva, and Diego Jarquin. Impact of environmental covariates summarization on predictive ability in genomic selection. *The Plant Genome*, 19(1):e70194, 2026.

Guangchao Sun, Jose Luis Araus, James C Schnable, Jie Xu, Henry T Nguyen, Yunbi Xu, and Yanli Lu. Enviromics and abiotic enviromics: Enhancing stress biology and plant resilience breeding. *Advanced Science*, pp. e77304, 2026.

George AK Van Voorn, Martin P Boer, Sandra Huynh Truong, Nicholas A Friedenberg, Shota Gugushvili, Ryan McCormick, Daniela Bustos Korts, Carlos D Messina, and Fred A Van Eeuwijk. A conceptual framework for the dynamic modeling of time-resolved phenotypes for sets of genotype-environment-management combinations: a model library. *Frontiers in Plant Science*, 14:1172359, 2023.

Kelin Wang, Muhammad Ali Abid, Awais Rasheed, Jose Crossa, Sarah Hearne, and Huihui Li. Dnngp, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, 16(1):279–293, 2023.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: a biological world model for virtual cell simulation. In *International Conference on Learning Representations*, volume 2026, pp. 72795–72823, 2026.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, pp. 22419–22430, 2021.

Zhou Yao, Mengting Yao, Chuang Wang, Ke Li, Junhao Guo, Yingjie Xiao, Jianbing Yan, and Jianxiao Liu. Geformer: A genotype-environment interaction-based genomic prediction method that integrates the gating multilayer perceptron and linear attention mechanisms. *Molecular Plant*, 18(3):527–549, 2025.

Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Ana Berthel, Zong-Yan Liu, Wei-Yun Lai, Zachary R Miller, Armin Scheben, Michelle C Stitzer, M Cinta Romay, et al. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained dna language model. *Proceedings of the National Academy of Sciences*, 122(24):e2421738122, 2025.

Dongfeng Zhang, Yanyun Han, Shouhui Pan, Zhongqiang Liu, Xiangyu Zhao, Qiusi Zhang, Qi Zhang, Xiaofeng Wang, Jiahao Sun, and Kaiyi Wang. Maizegep: A maize hybrids dataset with genotype, phenotype, and envirotype to develop genomic selection models. *Genomics, Proteomics & Bioinformatics*, pp. qzaf140, 2026.

Shuchang Zhou, Weipeng Fang, Runing Gao, Xi Long, Lu Chen, Xianliang Hu, and Ting Zhao. Gebiformer: bidirectional cross-attention integration of genomic and enviromic data for genotype-by-environment prediction in maize. *Briefings in Bioinformatics*, 27(4):bbag437, 2026.

A DATASETS

A.1 DATA PROCESSING

Each sample in GEWorld consists of three paired modalities—genotype, environment, and phenotype—with distinct structure and roles in the world model. We summarize the preprocessing and usage of each modality below; crop-specific sources are listed in Table S1.

Table S1: Summary of datasets used for static phenotype prediction and dynamic trajectory forecasting. Static datasets contain terminal genotype–environment observations, whereas dynamic datasets contain ordered weekly growth trajectories.

Protocol	Crop	Type	Genotypes	Markers	Environments	Samples	Trait	Reference
Static	Maize	Real	4,938	2,425	272	163,026	Yield	(Chen et al., 2026)
	Soybean	Real	1,379	4,611	6	8,274	Yield	(Sagae et al., 2026)
	Sorghum	Real	296	2,048	2	569	Dry biomass	(Burnette et al., 2018)
	Rice	Real	429	509	3	2,576	Plant height	(LaPorte et al., 2026)
	Wheat	Real	3,731	9,045	4	14,924	Yield	(Pérez-Rodríguez et al., 2017)
Dynamic	Sorghum	Simulated	1,500	1,000	4	6,000	Biomass, LAI	
	Wheat	Simulated	199	4,002	124	24,676	Biomass, LAI	(Bustos-Korts et al., 2019)
	Sorghum	Real	296	2,048	2	569	Height, cover	(Burnette et al., 2018)
	Wheat	Real	904	18,845	7	4,058	Height, cover	(Roth et al., 2025)

Genotype: Genotype provides a high-dimensional but *static* identifier for each individual: genome-wide SNP markers encoded as additive dosages in $\{0, 1, 2\}$ that remain fixed over the season. In GEWorld, genotype parameterizes variety identity and modulates G×E response. We project the marker vector through MLP_g to obtain the genotype embedding $\mathbf{z}_g$, which (i) participates in the gated G×E fusion used to initialize the latent state $\mathbf{s}_0$ and (ii) optionally conditions transition dynamics via genotype-dependent FiLM modulation during rollout. We mean-impute missing calls and z-score markers using statistics computed on the training split only. To standardize input dimensionality across methods, we retain the top-K loci by training-set variance.

Environment: Environment captures site-, season-, and weather-level drivers of development. Available covariates vary by dataset: maize and soybean include seasonal weather aggregates (e.g., temperature, radiation, precipitation); static sorghum provides growing-window summaries; rice and static wheat provide discrete location/trial identifiers; and dynamic sorghum and wheat provide weekly weather vectors indexed by days after sowing. For static phenotype prediction, we encode trial-level environment using either frozen MOMENT (Goswami et al., 2024) embeddings (maize, soybean, and dynamic-sorghum endpoints) or raw covariates/indicators (rice, static wheat, static sorghum). For dynamic rollouts, we always use the raw weekly weather sequence $\mathbf{e}_{1:T}$ as the exogenous input driving transitions.

Phenotype: Phenotype serves both as the supervised target and, in the dynamical setting, as the observation stream used for assimilation. Static tasks provide a single terminal trait per crop, decoded from the initialized state $\mathbf{s}_0$ without within-season measurements. Dynamic tasks provide weekly trajectories: the observed prefix is assimilated to update the latent state (transition + GRU correction), and post-cutoff observations supervise open-loop trajectory decoding under the known future environment. Thus, phenotypes are used both to train prediction heads and to correct the latent state before forecasting.

B EXPERIMENT DETAILS

B.1 METRICS

B.1.1 STATIC PHENOTYPE PREDICTION

All static metrics are computed on the held-out test split after mapping predictions back to the original trait scale. Let $\{(\mathbf{y}_i, \hat{\mathbf{y}}_i)\}_{i=1}^{n}$ denote the n finite ground-truth prediction pairs for a given crop and seed, and let $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$ be the mean of the observed targets.

R^2 **(coefficient of determination):** R^2 measures the fraction of phenotype variance explained by the model relative to predicting the constant mean $\bar{\mathbf{y}}$:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}} = 1 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})^2}, \tag{S1}$$

where $\mathrm{SS}_{\mathrm{res}}$ is the residual sum of squares and $\mathrm{SS}_{\mathrm{tot}}$ is the total sum of squares. $R^2 = 1$ indicates perfect reconstruction; $R^2 = 0$ matches the mean predictor; $R^2 < 0$ means worse than predicting $\bar{\mathbf{y}}$.

RMSE (root mean squared error): RMSE reports the absolute scale of prediction error in the same units as the trait:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}. \tag{S2}$$

Lower is better. Because RMSE retains the trait units, it is not directly comparable across crops with different scales.

PCC (Pearson correlation coefficient): PCC measures the linear association between predicted and observed phenotypes, independent of absolute offset and scale:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}})}{\sqrt{\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})^2}\sqrt{\sum_{i=1}^{n}(\hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}})^2}}, \tag{S3}$$

where $\bar{\hat{\mathbf{y}}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}_i$. PCC ranges in $[-1, 1]$; higher is better.

B.1.2 Dynamic trajectories evaluation

Dynamic scoring uses the same R^2, RMSE, and PCC definitions as the static endpoints above. The only difference is the evaluation set: given season length L and prefix ratio $\rho \in \{0.25, 0.50, 0.75\}$, the cutoff is $k = \max\big(1,\, \min(L-1,\, \lceil L \cdot \rho \rceil)\big)$. The model observes phenotypes (and environments) for $t \le k$, then rolls out open-loop under the known future environment sequence for $t > k$; future true phenotypes are never fed into the transition. Metrics are computed on all valid post-cutoff entries pooled over test trajectories, times $t > k$.

B.1.3 Counterfactual Intervention Evaluation:

Environment counterfactual evaluation asks whether swapping the future environment moves predictions in the same direction as the observed cross-environment phenotype change. For every genotype g observed under both environments e and e', define the observed and predicted directed deltas

$$\boldsymbol{\Delta}(g, e \to e') = \mathbf{y}(g, e') - \mathbf{y}(g, e), \qquad \hat{\boldsymbol{\Delta}}(g, e \to e') = \hat{\mathbf{y}}(g, e') - \hat{\mathbf{y}}(g, e). \tag{S4}$$

In the dynamic weather-swap setting, $\hat{\mathbf{y}}(g, e)$ and $\hat{\mathbf{y}}(g, e')$ denote the terminal phenotype predictions obtained from two open-loop rollouts that share the same latent state after observing the first 50% of the trajectory under environment e. The two rollouts use the same genotype, transition dynamics, and prediction heads, differing only in whether the future weather follows e or is replaced by e'. For each valid directed environment pair, we compute the observed response $\boldsymbol{\Delta}_i$ and the corresponding model response $\hat{\boldsymbol{\Delta}}_i$, yielding $\left\{(\boldsymbol{\Delta}_i, \hat{\boldsymbol{\Delta}}_i)\right\}_{i=1}^{N}$.

We report three trend metrics. Sign agreement measures directional consistency:

$$\mathrm{SignAgree} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[\mathrm{sign}(\boldsymbol{\Delta}_i) = \mathrm{sign}(\hat{\boldsymbol{\Delta}}_i)\big], \tag{S5}$$

where $\mathrm{sign}(0)$ is treated as a mismatch unless both sides are zero.

Pearson correlation on the deltas is

$$\mathrm{Pearson}_{\Delta} = \frac{\sum_{i=1}^{N}(\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}})(\hat{\boldsymbol{\Delta}}_i - \bar{\hat{\boldsymbol{\Delta}}})}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}})^2}\sqrt{\sum_{i=1}^{N}(\hat{\boldsymbol{\Delta}}_i - \bar{\hat{\boldsymbol{\Delta}}})^2}}. \tag{S6}$$

Spearman rank correlation replaces raw values by their ranks $r(\cdot)$:

$$\mathrm{Spearman}_{\Delta} = \frac{\sum_{i=1}^{N} \big(r(\boldsymbol{\Delta}_i) - \bar{r}_{\Delta}\big)\big(r(\hat{\boldsymbol{\Delta}}_i) - \bar{r}_{\hat{\Delta}}\big)}{\sqrt{\sum_{i=1}^{N} \big(r(\boldsymbol{\Delta}_i) - \bar{r}_{\Delta}\big)^2}\sqrt{\sum_{i=1}^{N} \big(r(\hat{\boldsymbol{\Delta}}_i) - \bar{r}_{\hat{\Delta}}\big)^2}}. \tag{S7}$$

All three metrics are higher-is-better; ties and constant vectors leave the correlation terms undefined.

B.2 Baselines

B.2.1 Static baselines

- **GBLUP**: classical genomic best linear unbiased prediction that models additive genetic effects via a genomic relationship or low-rank genomic basis together with environment main effects (Habier et al., 2007).
- **BGGE**: Bayesian genomic G×E package that builds multi-environment genomic covariance kernels and fits Bayesian linear mixed models for genomic-enabled prediction under G×E (Granato et al., 2018).
- **MoEGE2P**: mixture-of-experts genotype–envirotype–phenotype model from MaizeGEP: a softmax gating network routes concatenated SNP dosage and environmental covariates to multiple two-layer ReLU expert networks whose gated outputs are combined for phenotype prediction (Zhang et al., 2026).
- **KIBREED**: environmentally adapted genomic prediction framework that extends GBLUP-style models with growth-environment information for environment-specific performance, and uses CNN-based deep models over markers with an enviromic branch for G×E prediction (Gogna et al., 2026).
- **GEFormer**: genotype–environment interaction predictor that extracts SNP features with a gating MLP, models daily multi-factor weather with omni-dimensional dynamic convolution, captures temporal environmental dynamics with linear attention, and fuses genomic and environmental representations via a gating mechanism (Yao et al., 2025).
- **GE-BiFormer**: multimodal G×E framework with dual-path genomic and enviromic feature disentanglement, tokenized bidirectional cross-attention between SNP and environment tokens, and mixture-of-experts routing for phenotype prediction (Zhou et al., 2026).

All static methods share genotype-holdout splits, marker truncation, and continuous environment inputs under our unified protocol.

B.2.2 Dynamic baselines

- **Persistence**: seasonal-anomaly persistence of the last observed phenotype. The forecast holds the residual relative to the training seasonal mean and, when the climate curve declines after the cutoff, decays that residual toward zero so the trajectory follows the known seasonal shape (Hyndman et al., 2014).
- **HistoryMeanEnv**: ridge regression (α=1) that maps, at each future step t, the feature vector [last, prefix mean, seasonal$_t$, t_{rel}, weather$_t$] to the phenotype.
- **dynamicGP**: genotype-aware dynamical baseline adapted from Hobby et al. (Hobby et al., 2025). Per-genotype DMDc operators (A, B) are fit under known-future environment control; $\mathrm{vec}(A, B)$ are treated as multi-trait phenotypes and predicted for held-out genotypes by VanRaden-GRM GBLUP/GREML; forecasts assimilate the phenotype prefix then open-loop roll under the known future weather.
- **TCN**: temporal convolutional network under the shared prefix / known-future protocol. A causal TCN encodes the concatenated phenotype–environment prefix, and a future TCN decodes the remaining trajectory under the known future weather sequence; genotype enters as an additive conditioning vector on the prefix context (Lea et al., 2017).
- **Autoformer**: decomposition Transformer with Auto-Correlation. The encoder receives past phenotypes and past weather; the decoder is conditioned on known future weather in the decoder inputs, matching the same genotype-holdout and known-future evaluation protocol (Wu et al., 2021).

Table S2: Static GEWorld hyperparameters used in the five-crop genotype-holdout comparison. Sorghum denotes the TERRA-REF dry-biomass panel.

Crop	d	lr	batch	epochs/pat.	ens.	K	notes
Maize	512	2×10^{-4}	256	120 / 25	12	8192	raw genome; α=0.25
Soybean	256	5×10^{-4}	64	120 / 25	12	8192	raw genome+env; α=0.5
Sorghum	256	5×10^{-4}	32	100 / 20	7	2048	raw seasonal env
Rice	768	1×10^{-3}	32	120 / 25	10	8192	α=0.25
Wheat	512	5×10^{-5}	64	80 / 15	2	8192	default dropout

Table S3: Dynamic APSIM GEWorld hyperparameters.

Setting	Sorghum	Wheat
State width d	128	128
Learning rate	2×10^{-4}	2×10^{-4}
Batch size	16	16
Max epochs / patience	60 / 8	80 / 12
Weight decay	10^{-4}	10^{-4}
Max weekly steps	32	32
Prefix ratios	0.25/0.50/0.75	0.25/0.50/0.75
Genotype FiLM θ_g	on	on
Residual dynamic head	on	on

All dynamic baselines observe the identical phenotype/environment prefix and future weather sequence as GEWorld, and are scored only on post-cutoff steps.

B.3 Training hyperparameters

Table S2 lists per-crop static settings for GEWorld; Table S3 lists APSIM dynamical settings.

B.4 Algorithmic details

Algorithm S1 consolidates the procedures in Sections 3.2–3.4: gated G×E initialization of $\mathbf{s}_0$, closed-loop assimilation on the phenotype prefix $t \leq k$, open-loop rollout under a specified future environment for $t > k$, and (optionally) a paired counterfactual branch that reuses the same assimilated state. Static endpoint prediction reads out $\mathbf{s}_0$ only; dynamic forecasting and weather-swap interventions share the same transition and adaptive decoder.

C Results

C.1 Simulated dynamic trajectory rollout

Table S4 reports genotype-holdout trajectory forecasting on simulated APSIM Sorghum and Wheat under known future environments, using BiomassWt and LAI at three observed-prefix ratios.

C.2 Dynamic rollout curves

Figures S1 and S2 extend the representative 50%-prefix results in Figure 2 to all three prefix ratios. In both figures, rows correspond to 25%, 50%, and 75% observed prefixes, and columns distinguish datasets and traits.

Figure S3 shows the complementary known-environment weather-swap developmental curves for simulated Sorghum and Wheat, and real Sorghum and Wheat, including factual and counterfactual rollouts after swapping future weather between paired environments.

Algorithm S1 Forward pass of GEWorld

Require: Genotype $\mathbf{g}$, trial environment $\mathbf{e}$, environment sequence $\mathbf{e}_{1:T}$, time intervals $\Delta t_{1:T}$, cutoff k, prefix observations $\mathbf{y}^{\mathrm{dyn}}_{1:k}$, masks $\mathbf{m}_{1:k}$, and optional $\mathbf{e}'_{k+1:T}$

Ensure: Endpoint prediction, future trajectory, and optional intervention response

$\mathbf{z}_g \leftarrow \mathrm{MLP}_g(\mathbf{g})$, $\mathbf{z}_e \leftarrow \mathrm{MLP}_e(\mathbf{e})$
$\boldsymbol{\gamma} \leftarrow \sigma(\mathbf{W}[\mathbf{z}_g; \mathbf{z}_e] + \mathbf{b})$
$\mathbf{u} \leftarrow \mathrm{LayerNorm}(\boldsymbol{\gamma} \odot \mathbf{z}_g + (\mathbf{1} - \boldsymbol{\gamma}) \odot \mathbf{z}_e)$
$\widetilde{\mathbf{s}} \leftarrow \mathrm{MLP}_s([\mathbf{u}; \mathbf{z}_g \odot \mathbf{z}_e])$
$(\boldsymbol{\phi}_e, \boldsymbol{\psi}_e) \leftarrow \mathrm{Linear}(\mathbf{z}_e)$
$\mathbf{s}_0 \leftarrow \mathrm{LayerNorm}(\widetilde{\mathbf{s}} \odot [\mathbf{1} + \alpha \tanh(\boldsymbol{\phi}_e)] + \alpha \boldsymbol{\psi}_e)$
$(\boldsymbol{\mu}^{\mathrm{end}}, \boldsymbol{\ell}^{\mathrm{end}}) \leftarrow \mathrm{MLP}_{\mu\ell}(\mathbf{s}_0)$
$\mathbf{s} \leftarrow \mathbf{s}_0$, $\hat{\mathbf{y}} \leftarrow \mathbf{0}$
for $t = 1, \ldots, k$ **do**
  if $t = 1$ **then**
    $\bar{\mathbf{s}}_t \leftarrow \mathbf{s}_0$
  else
    $\bar{\mathbf{s}}_t \leftarrow \mathcal{T}_{\mathbf{z}_g}(\mathbf{s}, \mathbf{e}_t, \Delta t_t)$
  end if
  if $\mathbf{m}_t$ is valid **then**
    $\mathbf{s} \leftarrow \mathcal{A}(\bar{\mathbf{s}}_t, \mathbf{y}^{\mathrm{dyn}}_t, \mathbf{m}_t, \mathbf{e}_t)$
    $\hat{\mathbf{y}} \leftarrow \mathbf{y}^{\mathrm{dyn}}_t$
  else
    $\mathbf{s} \leftarrow \bar{\mathbf{s}}_t$
  end if
end for
$\mathbf{s}_k \leftarrow \mathbf{s}$
function ROLLOUT($\mathbf{s}_{\mathrm{init}}, \mathbf{e}_{a:b}, \Delta t_{a:b}, \hat{\mathbf{y}}_{\mathrm{prev}}$)
  $\mathbf{s} \leftarrow \mathbf{s}_{\mathrm{init}}$, $\hat{\mathbf{y}} \leftarrow \hat{\mathbf{y}}_{\mathrm{prev}}$
  for $t = a, \ldots, b$ **do**
    $\mathbf{s} \leftarrow \mathcal{T}_{\mathbf{z}_g}(\mathbf{s}, \mathbf{e}_t, \Delta t_t)$
    $\boldsymbol{\mu}^{\mathrm{res}}_t \leftarrow \hat{\mathbf{y}} + \mathrm{MLP}_{\mathrm{dyn}}(\mathbf{s})$
    $\mathbf{b}_t \leftarrow (1 - \eta)\hat{\mathbf{y}} + \eta \mathbf{c}_t$
    $\boldsymbol{\mu}^{\mathrm{ref}}_t \leftarrow \mathbf{b}_t + \tanh(\mathrm{MLP}_{\Delta}(\mathbf{s}))$
    $\boldsymbol{\rho} \leftarrow \sigma(\boldsymbol{\ell}_\rho)$
    $\boldsymbol{\mu}^{\mathrm{dyn}}_t \leftarrow (\mathbf{1} - \boldsymbol{\rho}) \odot \boldsymbol{\mu}^{\mathrm{res}}_t + \boldsymbol{\rho} \odot \boldsymbol{\mu}^{\mathrm{ref}}_t$
    $\hat{\mathbf{y}} \leftarrow \boldsymbol{\mu}^{\mathrm{dyn}}_t$
  end for
  return $\{\boldsymbol{\mu}^{\mathrm{dyn}}_t\}_{t=a}^{b}$
end function
$\hat{\mathbf{y}}^{\mathrm{f}}_{k+1:T} \leftarrow \mathrm{ROLLOUT}(\mathbf{s}_k, \mathbf{e}_{k+1:T}, \Delta t_{k+1:T}, \hat{\mathbf{y}})$
if $\mathbf{e}'_{k+1:T}$ is provided **then**
  $\hat{\mathbf{y}}^{\mathrm{cf}}_{k+1:T} \leftarrow \mathrm{ROLLOUT}(\mathbf{s}_k, \mathbf{e}'_{k+1:T}, \Delta t_{k+1:T}, \hat{\mathbf{y}})$
  $\hat{\boldsymbol{\Delta}}_{k+1:T} \leftarrow \hat{\mathbf{y}}^{\mathrm{cf}}_{k+1:T} - \hat{\mathbf{y}}^{\mathrm{f}}_{k+1:T}$
end if

C.3 ABLATION DETAILS

Figure 4 summarizes the main ablation study; Tables S5 and S6 report the corresponding seed-level statistics. Static ablations retrain the endpoint predictor after removing genotype, removing environment, or replacing gated multiplicative G×E fusion by concatenation, under the same genotype-holdout splits as Table 1.

Macro averages are means over the five crops of per-seed crop averages. Dynamic ablations retrain the Sorghum, while disabling phenotype assimilation, freezing the assimilated state without open-loop transition, or replacing adaptive decode with a residual-only head. We report canopy height and canopy cover R^2 at the 50% and 75% prefixes highlighted in Figure 4b. Across both settings,

Table S4: Dynamic trajectory forecasting under known future environmental conditions. Test R^2 is reported for post-cutoff predictions at 25%, 50%, and 75% observed-prefix lengths. Bold and underlined values indicate the best and second-best results in each row, respectively.

Crop	Trait	Prefix	Persistence	HistoryMeanEnv	dynamicGP	TCN	Autoformer	**GEWorld**
Sorghum (Simulated)	BiomassWt	25%	0.656±0.000	0.914±0.000	0.881±0.000	0.991±0.001	0.890±0.016	**1.000±0.000**
		50%	0.598±0.003	0.881±0.000	0.810±0.000	0.907±0.066	0.984±0.006	**0.999±0.000**
		75%	0.607±0.003	0.862±0.001	0.704±0.001	0.881±0.075	0.988±0.004	**0.999±0.000**
	LAI	25%	0.491±0.000	0.932±0.000	0.813±0.000	0.989±0.001	0.866±0.023	**0.998±0.001**
		50%	0.521±0.005	0.902±0.000	0.742±0.001	0.971±0.023	0.977±0.007	**0.997±0.000**
		75%	0.586±0.002	0.831±0.001	0.683±0.001	0.966±0.019	0.976±0.007	**0.995±0.001**
Wheat (Simulated)	BiomassWt	25%	0.701±0.005	0.701±0.005	0.849±0.008	0.976±0.011	0.900±0.034	**0.996±0.001**
		50%	0.579±0.009	0.580±0.009	0.854±0.006	**0.993±0.002**	0.968±0.003	**0.993±0.002**
		75%	0.820±0.006	0.826±0.006	0.925±0.004	**0.999±0.000**	0.979±0.012	0.998±0.001
	LAI	25%	0.233±0.015	0.233±0.015	0.622±0.006	0.229±0.016	0.832±0.040	**0.916±0.031**
		50%	0.225±0.014	0.225±0.014	0.714±0.011	0.961±0.012	0.952±0.009	**0.977±0.003**
		75%	0.611±0.014	0.611±0.014	0.688±0.016	**0.962±0.003**	0.929±0.028	**0.962±0.005**

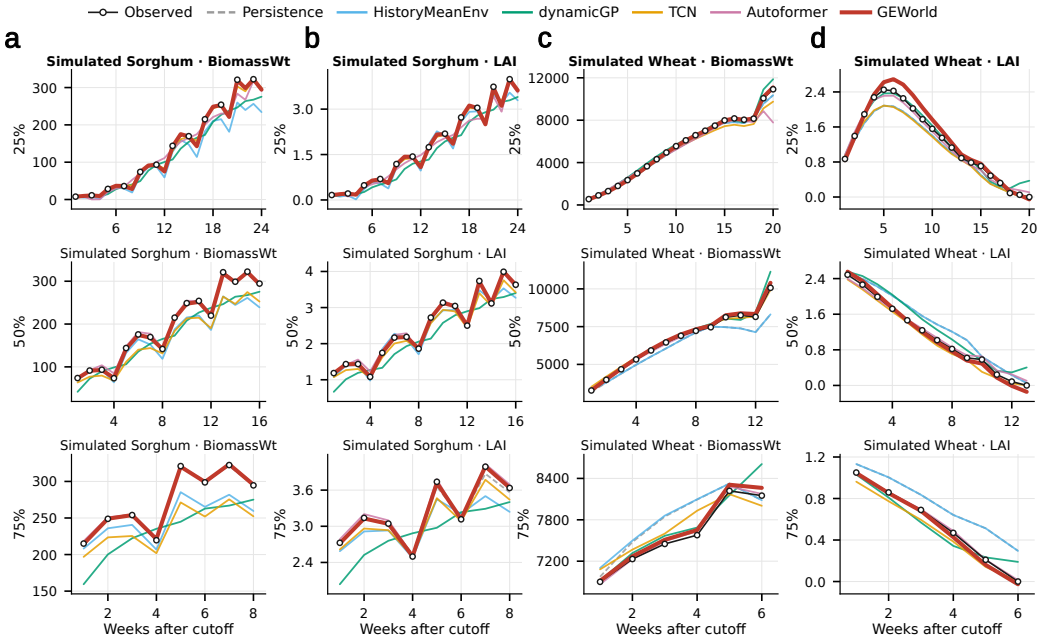


Figure S1: Dynamic forecasting on simulated Sorghum and Wheat. Rows correspond to 25%, 50%, and 75% observed prefixes; columns show BiomassWt and LAI for each crop. The horizontal axis denotes weeks after the forecast cutoff.

environment dominates most static panels, with Rice as the genotype-leaning exception, whereas the transition is the strongest dynamical necessity, and assimilation and decode matter increasingly for late-stage cover.

Table S5: Static input ablations. Variants are retrained without the indicated modality or G×E gating. Corresponds to Figure 4a.

Variant	Maize	Soybean	Sorghum	Rice	Wheat	Avg
GEWorld	0.661±0.006	0.761±0.008	0.560±0.090	0.574±0.076	0.958±0.002	0.703±0.028
w/o G×E	0.639±0.005	0.730±0.014	0.496±0.068	0.545±0.092	0.956±0.002	0.673±0.028
w/o genotype	0.574±0.018	0.653±0.018	0.362±0.060	0.018±0.016	0.949±0.003	0.511±0.016
w/o environment	0.094±0.006	0.052±0.014	0.143±0.034	0.526±0.088	0.005±0.001	0.164±0.016

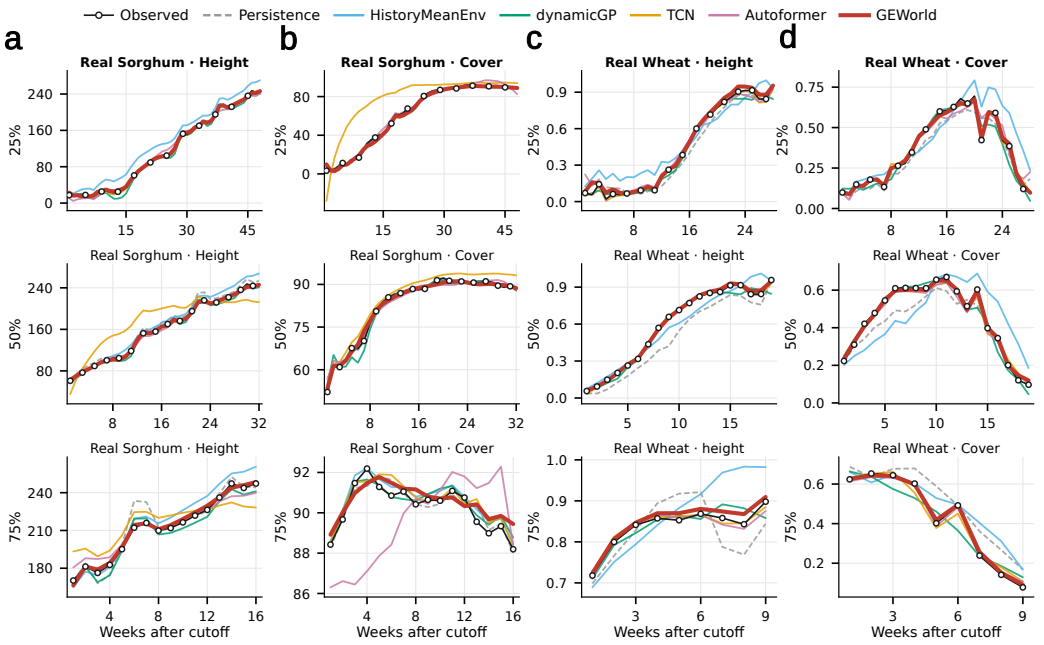


Figure S2: Dynamic forecasting on the real Sorghum and Wheat panels. Rows correspond to 25%, 50%, and 75% observed prefixes; columns show canopy height and canopy cover for each crop. The horizontal axis denotes weeks after the forecast cutoff.

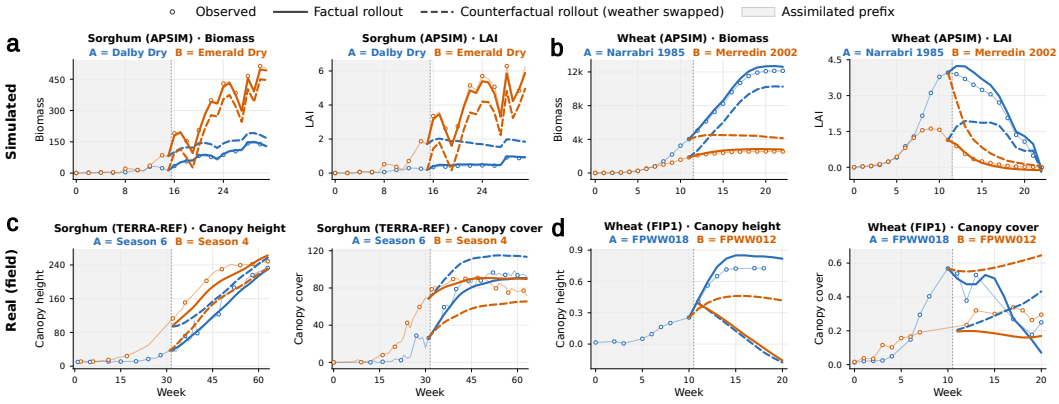


Figure S3: **(a)–(d)** show dynamic known-environment weather-swap developmental curves with a 50% prefix. The top row contains simulated Sorghum and Wheat, while the bottom row contains real Sorghum and Wheat. Each panel combines mirrored $A \leftrightarrow B$ swaps for environment. Markers denote observed trajectories, with colour indicating environment. Solid and dashed curves denote factual and counterfactual rollouts, with colour indicating future weather. Shaded regions mark the assimilated prefix.

C.4 Embedding feature analysis

A central design choice in GEWorld is the gated G×E fusion that forms s_0. We investigate whether this gate allocates attention to genotype and environment in a species-dependent manner, and whether such allocation reflects predictive necessity. Using trained static checkpoints under genotype holdout, we compute the test-set gate value $\gamma = \sigma(W[z_g; z_e])$ (where $\gamma \to 1$ favors genotype and $\gamma \to 0$ favors environment), and quantify each modality's contribution via inference-time zeroing of z_g, measured as ΔR^2 relative to the intact model (Figure S4).

Table S6: Dynamic module ablations on TERRA-REF Sorghum. Corresponds to Figure 4b.

Variant	Height@50%	Height@75%	Cover@50%	Cover@75%
GEWorld	0.920±0.013	0.780±0.028	0.776±0.035	0.502±0.087
w/o Assimilation	0.912±0.003	0.676±0.027	0.745±0.019	0.340±0.078
w/o Transition	0.216±0.023	0.450±0.026	0.018±0.029	0.361±0.109
w/o Decode	0.906±0.017	0.763±0.024	0.713±0.057	0.413±0.171

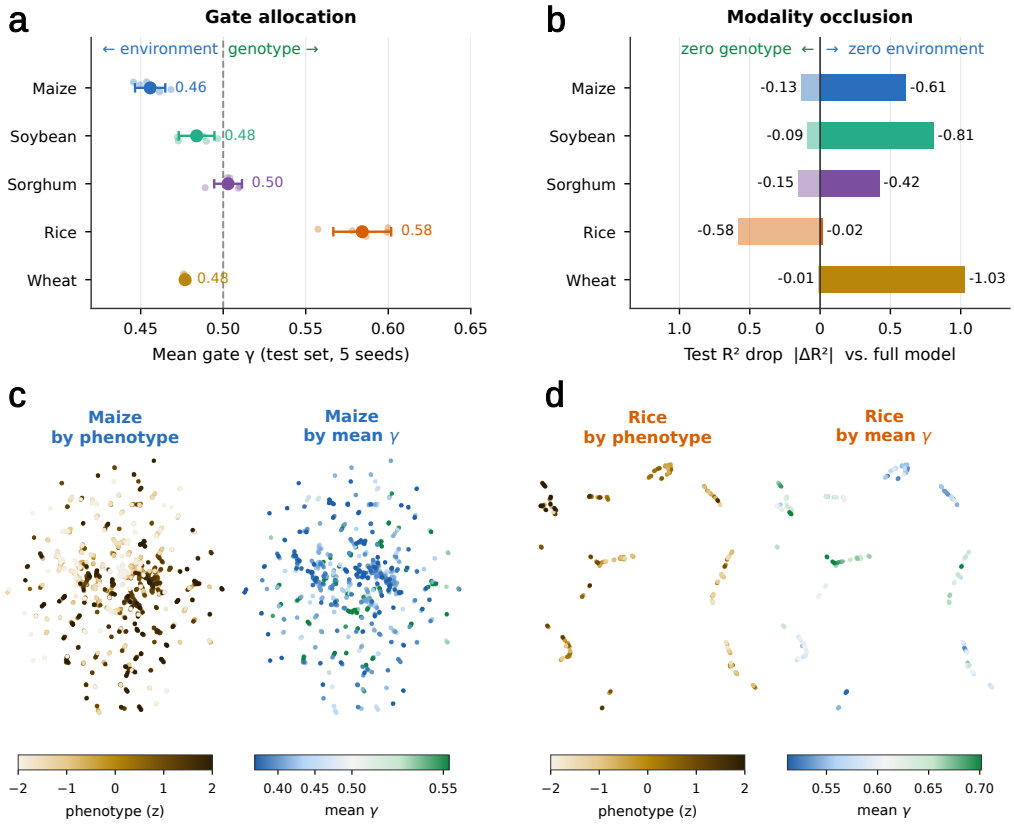


Figure S4: **(a)** Distribution of mean gate γ across crops (violins over seeds; dashed line at 0.5; $\gamma \to 1$ favors genotype, $\gamma \to 0$ favors environment). **(b)** Change in test R^2 under modality occlusion (zero genotype vs. zero environment). **(c)** UMAP of the initialized state $\mathbf{s}_0$ for Maize, colored by phenotype (z-scored) or mean γ. **(d)** Corresponding UMAP for Rice under the same colorings.

The gate is not a fixed half–half mix. Rice alone tilts toward genotype, while Maize, Soybean, Sorghum, and Wheat remain near the midpoint (Figure S4a). Occlusion recovers a complementary taxonomy (Figure S4b). Suppressing continuous environment sharply degrades Wheat and Soybean and remains costly for Maize and Sorghum, while suppressing genotype leaves Wheat nearly unchanged and only mildly hurts Soybean. Rice exhibits the opposite pattern: removing genotype costs R^2, whereas removing environment costs only -0.02, consistent with limited continuous enviromic signal on that panel.

Two-dimensional projections of s_0 (Figure S4c–d) exhibit coherent structure under both phenotype and gate colorings, indicating that the fused state preserves organized variation rather than collapsing to unstructured noise. Together, these analyses support the view that gated fusion adapts along a genotype–environment axis in a data-dependent way; they complement the retrained input ablations in Table S5.

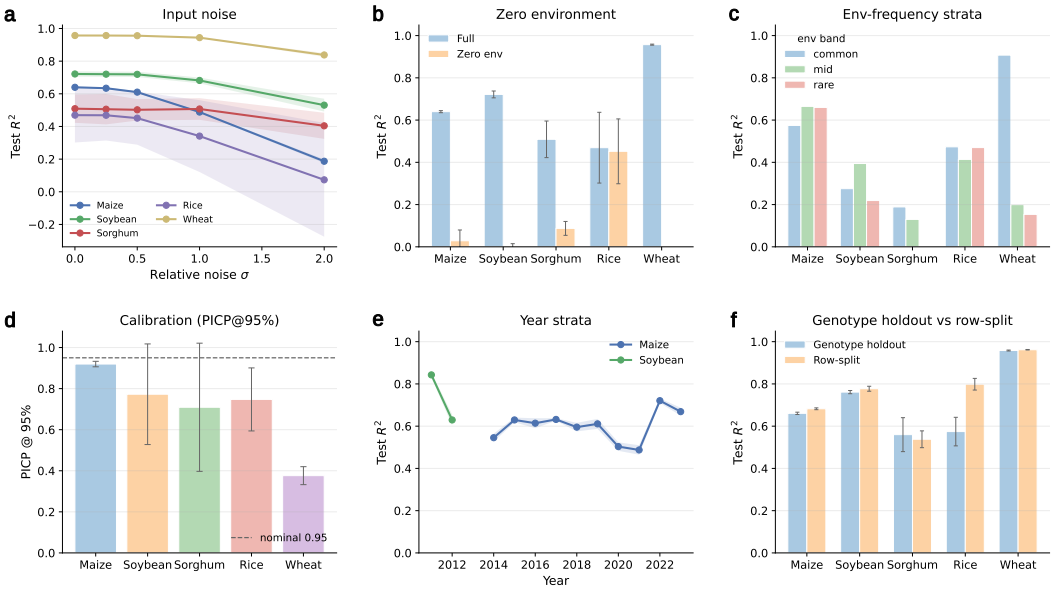


Figure S5: Robustness of static phenotype prediction across five crops. Top left to bottom right: input noise; drop environment-ID; environment-frequency strata; calibration (PICP@95%); year strata (Maize and Soybean); genotype holdout vs. phenotype-row random splits.

C.5 Robustness analysis

Beyond the genotype-holdout leaderboard (Table 1), we ask how stable static GEWorld predictions remain under corrupted inputs, missing environment identity, uneven environment coverage, season shifts, and an alternative data partition. Figure S5 summarizes these checks on the five static datasets.

Joint corruption of genotype and environment inputs yields a clear, monotonic decline in test R^2 on every crop (Figure S5a). At relative noise σ=2, Rice and Maize fall furthest from their clean scores, while Wheat remains comparatively resilient; Sorghum and Soybean degrade more gradually. Zeroing continuous environment features at inference produces a crop-dependent drop that mirrors the embedding occlusion analysis (Figure S5b and Figure S4b): Wheat and Soybean collapse to negative R^2, Maize and Sorghum fall near zero, and Rice changes little, indicating limited continuous enviromic signal on that panel.

Stratifying test environments by training frequency further separates easy and hard contexts (Figure S5c). Maize stays relatively flat across common, mid, and rare sites, whereas Wheat deteriorates sharply off the common band and Soybean remains weak on both common and rare strata. Uncertainty quality does not track point estimates one-to-one (Figure S5d). Nominal 95% intervals are near well calibrated on Maize but under-covered on Sorghum and especially Wheat ($\approx$0.38). Season-resolved scores echo this unevenness at the calendar scale (Figure S5e). Maize remains usable across 2014–2023 with noticeable year-to-year variation, while Soybean's two available seasons already differ substantially.

Finally, we examined the sensitivity of the model to the partitioning rules themselves. Figure S5f compares the genotype-holdout scheme with the random partition used by GEWorld; in the latter, the partitioning of rows is independent of genotype IDs, so the same genotype may appear in both the training and test sets. When the environmental mean is predictable, this simpler partitioning scheme may lead to inflated absolute R^2 values, yet it remains a standard approach for assessing the robustness of G$\times$E (genotype-by-environment interaction). Therefore, we re-evaluated all static baseline models under the condition of random partitioning.

Table S7 reports that GEWorld retains the best macro average (0.752) and leads on Maize, Soybean, Sorghum, and Wheat; GEFormer remains slightly ahead on Rice. On Sorghum, the row-split score

stays close to the genotype-holdout mean, while Rice rises sharply under row leakage; the ranking among deep models remains broadly stable.

Table S7: Static phenotype prediction test R^2 under phenotype-row random $70/15/15$ splits, with higher values indicating better performance. The best result in each column is shown in bold.

Method	Maize	Soybean	Sorghum	Rice	Wheat	Avg
GBLUP	0.599 ± 0.006	0.684 ± 0.010	0.483 ± 0.076	0.627 ± 0.046	0.950 ± 0.001	0.669
BGGE	0.615 ± 0.007	0.689 ± 0.011	0.400 ± 0.106	0.575 ± 0.078	0.951 ± 0.001	0.646
MoEGE2P	0.607 ± 0.006	0.286 ± 0.016	0.413 ± 0.091	0.750 ± 0.040	0.941 ± 0.002	0.599
KIBREED	0.657 ± 0.007	0.333 ± 0.016	0.449 ± 0.088	0.753 ± 0.051	0.952 ± 0.001	0.629
GEFormer	0.669 ± 0.006	0.711 ± 0.012	0.500 ± 0.025	$\mathbf{0.819\pm0.021}$	0.953 ± 0.001	0.731
GE-BiFormer	0.422 ± 0.157	0.325 ± 0.020	0.328 ± 0.172	0.729 ± 0.058	0.756 ± 0.428	0.512
GEWorld	$\mathbf{0.682\pm0.005}$	$\mathbf{0.777\pm0.013}$	$\mathbf{0.538\pm0.044}$	0.799 ± 0.031	$\mathbf{0.962\pm0.001}$	**0.752**

C.6 Dynamic environment intervention details

Table S8 evaluates whether predicted responses to future weather swaps agree with observed cross-environment changes after a shared 50% phenotype prefix. Agreement is high on simulated APSIM data, remains positive on real Sorghum, and is weaker on real Wheat, indicating that counterfactual response quality depends on the data regime and should not be interpreted as uniform across panels.

Table S8: Dynamic known-environment weather-swap trend agreement at a 50% prefix. For the simulated APSIM panels, agreement is measured between the predicted and observed cross-environment changes in terminal BiomassWt. For the real TERRA-REF Sorghum and FIP1 Wheat panels, the corresponding trait is terminal canopy cover. Higher values indicate better agreement.

Panel	Species	Sign agree ↑	Spearman($\hat{\Delta}, \Delta$) ↑	Pearson($\hat{\Delta}, \Delta$) ↑
Simulated	Sorghum(APSIM)	1.000 ± 0.000	0.994 ± 0.003	0.996 ± 0.002
	Wheat(APSIM)	0.904 ± 0.006	0.944 ± 0.007	0.945 ± 0.007
Real	Sorghum (TERRA-REF)	0.898 ± 0.033	0.783 ± 0.044	0.707 ± 0.035
	Wheat (FIP1)	0.596 ± 0.067	0.380 ± 0.165	0.321 ± 0.115