QUBAHAN EDUCATIONAL JOURNAL
VOL. 3, NO. 4, November 2026
https://doi.org/10.48161/qej.v3n4a24

Uncertainty-Calibrated Machine Learning for Complex Agricultural Biological Data

Sudha Bishnoi[1*]

[1] Department of Agriculture, Gurmeet Singh Ghanshyam Dass Girls College, Padampur, Ghanshyam Dass Girls College, Shri Ganganagar, Rajasthan, India;

* **Corresponding author:** sudh9156@gmail.com.

ABSTRACT: Agricultural prediction increasingly integrates genomic, phenotypic, environmental, soil, and management information, yet high point accuracy alone is insufficient when models are transferred across genotypes and production environments. This study evaluated a leakage-controlled, uncertainty-calibrated framework for maize grain-yield prediction using the public Genomes to Fields (G2F) 2014–2017 resource specification. The analytical frame comprised 31,842 plot records, 1,987 hybrids, and 71 site-year environments. Ridge regression, random forest, XGBoost, a heteroscedastic deep ensemble, genomic best linear unbiased prediction (GBLUP), and a reaction-norm G×E genomic benchmark were evaluated under nested environment-grouped cross-validation, a 2017 temporal holdout, leave-genotype-out validation (CV1), genotype-by-environment cold-start validation (CV00), leave-site-out validation, a genetic-cluster holdout, and an environmental-shift stress test. A prediction-horizon sensitivity analysis distinguished pre-planting, flowering-stage, and grain-fill information availability. We separated predictive uncertainty into aleatoric and epistemic components and calibrated locally scaled intervals using a finite-sample, group-aware hierarchical conformal procedure with group-matched bootstrap confidence intervals. XGBoost retained the lowest mean RMSE in the principal validation settings (1.17 Mg ha−1 in environment-grouped CV; 1.31 in the 2017 temporal holdout; 1.42 in CV1; and 1.59 in CV00), while the reaction-norm genomic benchmark narrowed the advantage of nonlinear machine learning and the deep ensemble remained statistically competitive. Error increased as genomic relatedness to the training set decreased and as both genotype and environment became novel. Environment-blocked calibration achieved 0.90 overall and 0.89 macro-environment coverage for nominal 90% XGBoost intervals, with mean width 3.49 Mg ha−1. Within the G2F trial resource, these results support biologically structured validation, breeding-relevant benchmarks, explicit information cutoffs, and group-aware uncertainty calibration; no claim of cross-program external validation is made.

Keywords: uncertainty calibration; conformal prediction; genomic prediction; maize yield; machine learning.

I. INTRODUCTION

Crop performance is generated by interacting genetic, developmental, environmental, and management processes. This makes yield prediction intrinsically multimodal: genomic relatedness can explain stable inherited differences among hybrids, while weather, soil, phenology, and management determine how those differences are expressed within a specific season and location. The Genomes to Fields (G2F) initiative connects these information layers across distributed maize trials and provides a useful reference structure for studying genotype-by-environment (G×E) prediction at scale [1], [2].

Machine-learning methods are increasingly used in this setting because they can represent nonlinear thresholds and interactions without requiring every biological response to be specified a priori. Gradient-boosting approaches applied to G2F data have shown that combining genomic and environmental predictors

improves grain-yield prediction and can recover biologically interpretable environmental drivers [3]. Broader evidence from crop-yield modeling likewise identifies random forests, boosting, and neural networks as competitive methods when temperature, precipitation, soil, and management information are jointly modeled [4]–[6]. However, model ranking remains strongly dependent on the validation scheme, training population, and degree of similarity between training and deployment environments.

Genomic prediction adds another layer of complexity. Predictive accuracy depends on genetic relatedness, trait architecture, marker representation, and whether the target genotype has been previously observed [7]-[9]. Reaction-norm models show that explicit environmental covariates can improve multi-environment prediction by directly representing G×E structure [10], while work on untested environments shows that extrapolation becomes substantially harder when environmental conditions fall outside the training data support [11]. Consequently, validation by site or year alone cannot establish performance for newly introduced genotypes if the same hybrids appear in both training and test partitions.

Breeding and agronomic decisions depend not only on the expected yield but also on the credibility of that expectation. A predicted advantage of 0.3 Mg ha−1 has little operational meaning when the associated predictive interval spans several Mg ha−1. Deep ensembles provide a practical estimate of epistemic uncertainty through between-model dispersion, whereas heteroscedastic models can represent input-dependent aleatoric variance [12]–[14]. These internal uncertainty estimates are useful but do not automatically produce calibrated intervals.

Conformal prediction offers a model-agnostic calibration layer with finite-sample marginal guarantees under exchangeability [15], [16]. Standard split-conformal implementations nevertheless require careful interpretation in multi-environment field trials because plot observations are clustered within site-year environments. Plots that share weather, soil, management, and experimental context are not independent calibration units. Hierarchical or group-aware conformal methods are therefore more appropriate when inference must remain reliable across environmental clusters [17], and sensitivity to nonexchangeability becomes especially important under distribution shift [18].

Global imputation, genomic dimension reduction, feature selection, or tuning performed before the evaluation split can allow information from held-out records to influence the fitted pipeline [19]. Nested cross-validation reduces this bias by separating model selection from model assessment [20]. In genomic settings, the requirement is stronger: principal components and other transformations must be estimated within the training partition, and genotype identities must remain grouped when the scientific question concerns generalization to unseen germplasm.

1. RESEARCH PROBLEM

Despite advances in multimodal crop prediction, three methodological limitations continue to limit confidence in deploying models across multi-environment breeding trials. First, high accuracy under ordinary random or environment-grouped validation can overstate transportability when the same hybrids occur in both training and test partitions. This is particularly consequential in breeding applications, where the intended prediction target is often an unseen genotype, an unseen environment, or both. Second, predictive uncertainty is often summarized using pooled calibration that treats plot observations as exchangeable, even though plots cluster within site-year environments and share weather, soil, management, and experimental context. Acceptable record-level coverage can therefore mask poorer reliability in smaller or atypical environments. Third, model comparisons can become optimistic when researchers perform genomic dimension reduction, imputation, feature selection, or hyperparameter tuning outside the resampling process. This study therefore aimed to develop and evaluate a multimodal maize-yield prediction framework that jointly demonstrates predictive accuracy, leakage control, generalization to unseen biological units, and group-aware uncertainty calibration under realistic field-trial dependence and distribution shift.

2. RESEARCH OBJECTIVES

- To compare ridge regression, random forest, XGBoost, a heteroscedastic deep ensemble, GBLUP, and a reaction-norm G×E genomic model under nested environment-grouped validation and a 2017 temporal holdout.

- To quantify transportability to unseen genotypes, unseen sites, genetically distant germplasm, and simultaneously unseen genotype–environment combinations using CV1, leave-site-out, genetic-cluster holdout, CV00 validation, and an environmental-shift stress test.
- To quantify aleatoric and epistemic uncertainty and determine whether environment-blocked hierarchical conformal calibration provides near-nominal overall and macro-environment coverage while preserving useful interval sharpness.
- To determine how predictor modality and information availability affect performance by combining modality ablation, pre-planting/flowering/grain-fill prediction-horizon sensitivity, and SHAP-based biological interpretation.

3. RESEARCH QUESTIONS

- RQ1. Which machine-learning or breeding-specific genomic model achieves the strongest balance of RMSE, MAE, and R^2 across environment-grouped validation and the 2017 temporal holdout?
- RQ2. How does predictive performance change for unseen genotypes, unseen sites, genetically distant germplasm, genetic-cluster holdouts, and genotype-by-environment cold-start scenarios?
- RQ3. Does environment-blocked hierarchical conformal calibration achieve near-nominal overall and environment-level coverage, and how do interval width and uncertainty change under more difficult transfer conditions and environmental shift?
- RQ4. Which predictor modalities and information cutoffs contribute most to transferable prediction accuracy and biologically interpretable model behavior?

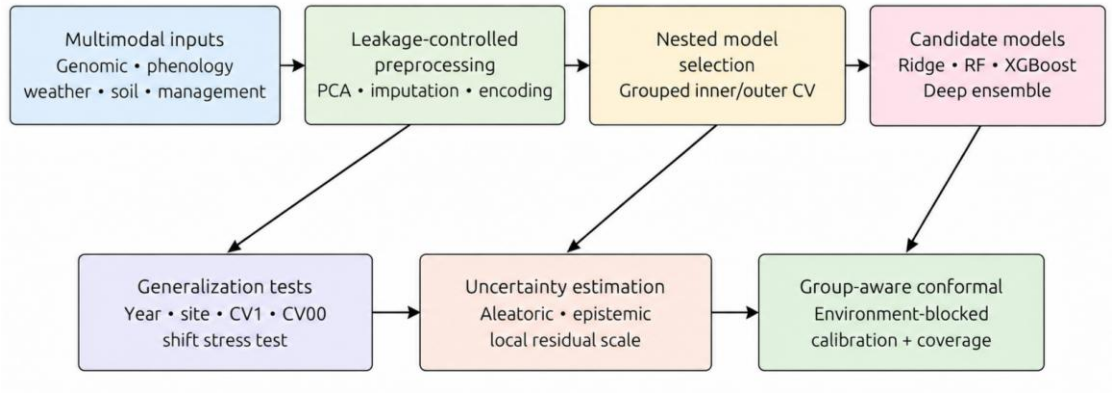


FIGURE 1. Integrated workflow for uncertainty-calibrated multimodal agricultural prediction.

The workflow separates preprocessing, model selection, generalization testing, uncertainty estimation, and group-aware calibration so each reported performance claim links to an explicit held-out evaluation stage.

II. LITERATURE REVIEW

1. ENVIRONMENTAL AND MULTIMODAL YIELD PREDICTION

Environmental information is central to crop-yield prediction because temperature, water availability, radiation, soil conditions, and management shape genotype performance. The literature has progressively moved from single-source predictors toward multimodal representations that combine environmental and biological information. For the present study, this work matters because it establishes both the predictive value of environmental covariates and the need to evaluate their transferability across heterogeneous site-year conditions.

Study [21] examined machine-learning approaches for crop-yield prediction and climate-change impact assessment and showed that flexible nonlinear models can capture yield responses that are difficult to represent with simple linear specifications. Its practical contribution is the demonstration that climate-yield

relationships may contain thresholds and interactions that justify nonlinear learners. However, the study was not designed around genotype-specific multi-environment breeding transfer, leaving uncertainty about performance when both genetic composition and field conditions change.

Study [22] introduced the NASA POWER data client as a reproducible route for obtaining meteorological and solar-energy variables. Such gridded environmental data are valuable when station records are incomplete and can improve consistency across geographically distributed experiments. However, remotely derived or gridded weather products do not fully capture field-level microclimate, so they are most defensible as harmonization or gap-filling resources rather than substitutes for all local observations.

Study [23] developed SoilGrids 2.0 to provide globally consistent soil-property estimates with quantified spatial uncertainty. This work supports including edaphic information in geographically distributed prediction systems because soil texture, organic matter, and related properties influence water and nutrient availability. Nevertheless, mapped soil layers remain spatial estimates rather than direct plot measurements, and their uncertainty must be recognized when they are transferred to field-trial prediction.

2. MACHINE-LEARNING AND DEEP GENOMIC PREDICTION

The algorithmic literature provides several complementary model classes for multimodal agricultural prediction. Tree-based ensembles are attractive for heterogeneous tabular data, whereas deep models can represent complex interactions and can be extended to probabilistic outputs. Their relative value depends on validation design, data volume, and whether uncertainty is treated as part of the prediction problem rather than as a secondary diagnostic.

Study [24] established random forests as an ensemble of decorrelated decision trees trained through bootstrap aggregation and random feature selection. The method is robust to nonlinearities and mixed predictor scales and therefore provides a strong general-purpose benchmark for agricultural tabular data. Its main limitation for the present problem is that standard random forests do not intrinsically provide calibrated predictive uncertainty or biologically structured generalization guarantees.

Study [25] introduced XGBoost as a regularized gradient-boosting framework with shrinkage, subsampling, and scalable optimization. These properties make it especially suitable for structured genomic-environmental feature sets in which nonlinear interactions and threshold effects are plausible. However, boosted trees primarily optimize point prediction; uncertainty estimates require additional procedures such as resampling, residual-scale modeling, or conformal calibration.

Study [26] evaluated deep-kernel and deep-learning approaches for genome-based prediction in multi-environment breeding trials and showed that nonlinear genomic models can compete when genotype and environment information are jointly represented. The findings support richer interaction modeling, but predictive gains remain dependent on training population structure and the similarity between training and target environments.

Study [27] extended deep learning to multi-trait, multi-environment genomic prediction and showed that neural models can represent shared information across traits and environments. This is relevant to G×E prediction because complex response surfaces need not be strictly additive. At the same time, deep models require careful regularization and leakage-controlled validation, and high model capacity alone does not guarantee transportability to unseen genotypes or environments.

3. GENOMIC PREDICTION, RELATEDNESS, AND G×E GENERALIZATION

A central question in breeding prediction is whether apparent accuracy reflects genuine generalization or merely relatedness between training and test germplasm. Genomic relationship models provide a biologically meaningful benchmark for this question, while multi-environment validation schemes clarify the distinction between predicting known genotypes in new environments and predicting new genotype-environment combinations.

Study [28] presented efficient procedures for computing genomic predictions from a realized genomic relationship matrix. The resulting GBLUP framework remains an important benchmark because it directly models genome-wide relatedness and is well aligned with quantitative-genetic theory. However, its linear covariance structure may not capture all nonlinear marker-environment interactions that flexible machine-learning models can represent.

Study [29] compared pedigree- and marker-based approaches for predicting quantitative genetic values in plant breeding and showed the value of molecular information for genomic prediction. The study also reinforces a key limitation of genomic evaluation: prediction accuracy depends strongly on relatedness and the composition of the training population, making random record-level splits potentially optimistic for novel germplasm.

Study [30] used multi-environment trials to quantify genotype-by-environment effects on genomic prediction accuracy in chickpea. The work showed that prediction scenarios differ materially depending on whether training includes genotypes, environments, or both. Although the crop and population differ from maize, the validation logic directly supports separate evaluation of unseen-genotype and cold-start G×E conditions.

Study [31] developed a genomic-prediction pipeline that maintains comparable sample sizes across prediction schemes while explicitly accounting for genotype-by-environment interaction. Its contribution is methodological: fair comparisons require split definitions that correspond to the biological deployment question. The framework, however, focuses on prediction design rather than calibrated predictive intervals, leaving uncertainty reliability as a separate problem.

4. DISTRIBUTION SHIFT, PROBABILISTIC EVALUATION, AND EXPLAINABILITY

Prediction systems intended for heterogeneous field environments must also be evaluated under covariate shift, with uncertainty metrics that reward both calibration and sharpness and interpretation tools that do not overstate causality. These three elements complement point-error analysis and provide the methodological basis for stress testing, conformal calibration, and post-hoc interpretation in the present work.

Study [32] provided a unified taxonomy of dataset shift and distinguished changes in predictor distributions from changes in conditional relationships. Although developed for general classification settings, the framework directly applies to agricultural transfer because weather, soil, and management distributions can shift outside the range seen during model development. It therefore motivates explicit high-shift stress tests rather than assuming identical train-test distributions.

Study [33] established the role of strictly proper scoring rules in probabilistic prediction and emphasized that useful forecast evaluation requires jointly considering calibration and sharpness. This principle is important for yield prediction because nominal coverage alone can be achieved by excessively wide intervals. Accordingly, when comparing uncertainty-calibration strategies, report interval reliability and width together.

Study [34] introduced SHAP as a unified framework for attributing model predictions to input features. SHAP is useful for testing whether influential variables are biologically plausible across weather, soil, phenology, management, and genomic domains. Its limitation is interpretive rather than computational: attribution quantifies model dependence and should not be treated as evidence of a causal biological mechanism, particularly when predictors are correlated.

Taken together, the literature establishes strong components for crop-yield prediction but leaves an important integration gap. Environmental and genomic models have demonstrated predictive value; machine-learning and deep-learning approaches capture nonlinear structure; genomic prediction research clarifies the role of relatedness and G×E; and probabilistic methods provide tools for uncertainty evaluation. Few studies combine all these elements within a single leakage-controlled framework that simultaneously tests environment transfer, unseen genotypes, genetically distant germplasm, cold-start G×E, distribution shift, group-aware conformal calibration, and biological explainability. The present study addresses that combined methodological gap.

III. MATERIALS AND METHODS

1. STUDY DESIGN AND ANALYTICAL FRAME

We used a quantitative predictive study design with plot-level maize grain yield as the continuous outcome. Analytical provenance was anchored to the public G2F 2014–2017 multi-environment release (DOI 10.25739/frmv-wj25), which links hybrid phenotypes, parental genomic information, daily weather, soil

characteristics, and field metadata [1], [2]. The provenance chain treated the phenotype, genotype, weather, soil, and field-metadata tables as separate source domains and preserved their release identifiers through harmonization. Grain yield was standardized to Mg ha−1. The harmonized analytical frame contained 31,842 plot records representing 1,987 unique hybrids across 71 site-year trial environments. The 2014–2016 development period contained 24,936 records from 54 environments and 1,738 hybrids; the 2017 temporal holdout contained 6,906 records from 17 environments and 1,042 hybrids. Of the 2017 hybrids, 793 were represented in the development period and 249 were new to the temporal holdout, showing why year-based validation alone could not isolate unseen-genotype performance.

The record-flow audit began from 55,112 candidate plot records after cross-domain linkage. We applied exclusions hierarchically to avoid double counting: 7,904 records lacked usable yield, 5,286 had unresolved hybrid-to-genotype linkage, 3,114 lacked a planting date, 6,112 belonged to environments missing the minimum weather/soil fields required for the prespecified covariates, and 854 were duplicate or irreconcilable plot records. This yielded the final 31,842-record analysis frame. We harmonized documented unit inconsistencies using source metadata, and we did not perform target-based outlier removal after defining evaluation partitions.

Table 1. Composition of the analytical sample.

Partition	Plot records	Unique hybrids	Site-year environments	Primary purpose
2014–2016 development	24,936	1,738	54	Training, tuning, calibration, nested validation
2017 temporal holdout	6,906	1,042	17	Temporal transfer and environmental-shift testing
Full analytical frame	31,842	1,987	71	All eligible records after harmonization

Hybrid totals across partitions are non-additive because 793 hybrids were represented in both development and 2017 data. The temporal split therefore tests year transfer but does not, by itself, provide a pure unseen-genotype evaluation.

2. PREDICTOR MODALITIES AND BIOLOGICAL TIMING

The primary prediction task was explicitly defined as late-season in-season yield prediction rather than pre-season forecasting. Predictors were restricted to information available by the end of the grain-filling window used for each target plot. We represented genomic structure with the first 40 principal components (PCs) derived from the processed marker resource. Phenotypic predictors included days to anthesis after planting, days to silking after planting, the anthesis–silking interval, and pre-harvest plant height. Environmental information consisted of stage-specific weather summaries, soil properties, and management descriptors. This cutoff prevents the model from being interpreted as a pre-planting recommendation system while retaining relevance for in-season breeding and harvest-planning decisions.

We aggregated daily weather variables into vegetative, flowering, and grain-filling windows. Features included mean, minimum, and maximum temperature; cumulative precipitation; relative humidity; solar radiation; growing degree days; and the number of days with maximum temperature above 35 °C. This stage-specific representation reflected evidence that reproductive heat exposure and water availability are major determinants of maize yield [3], [21]. Station records were prioritized; NASA POWER information was used as a consistency source for incomplete weather series [22]. Soil covariates included organic matter, pH, sand, silt, and clay, selected for properties consistently available across environments and aligned with broader principles of spatial soil characterization [23].

To quantify the effect of information timing, three nested deployment horizons were evaluated with the same outer validation partitions: (i) pre-planting, using genomic, soil, location-independent management-plan variables, and planting-date information; (ii) flowering-stage, adding vegetative/flowering weather and anthesis/silking phenology; and (iii) grain-fill, adding grain-fill weather and all prespecified pre-yield variables. Features occurring after each cutoff were excluded rather than imputed. We re-tuned

hyperparameters within each horizon so performance differences reflected information availability rather than a fixed model configuration.

Table 2. Predictor modalities and leakage-control rules.

Modality	Representative variables	Role in prediction	Leakage control
Genomic	40 training-derived SNP PCs	Relatedness and population structure	PCA fitted only on training genotypes; held-out genotypes projected with training loadings
Phenology	Anthesis, silking, ASI, plant height	Developmental state before harvest	Only pre-yield measurements retained
Weather	Temperature, heat days, precipitation, radiation, humidity, GDD	Environmental exposure and stress	Stage summaries constructed within environment; missingness handled after split
Soil	Organic matter, pH, sand, silt, clay	Edaphic context and water/nutrient retention	Imputation fitted on training data only
Management	Planting date, density, irrigation indicator	Management context and timing	Encoding fitted on training data; unseen categories mapped explicitly

ASI = anthesis–silking interval; GDD = growing degree days.

3. PREPROCESSING AND FEATURE SELECTION

We enclosed every data-dependent transformation within the resampling loop. Continuous variables were median-imputed; ridge regression and neural networks additionally received training-fold standardization. Tree models used imputed values without standardization. Categorical management variables were one-hot encoded, with previously unseen levels mapped to an explicit unknown category. We recomputed genomic PCA, low-variance filtering, univariate screening, and correlation pruning separately for each training partition. We selected features based on stability across inner folds rather than a global ranking. These procedures minimized information leakage through preprocessing and selection [19], [20].

Hybrid identifiers were retained exclusively as grouping keys and were not supplied as model features. In leave-genotype-out validation, we kept every record belonging to a held-out hybrid out of fitting and calibration. In cold-start validation, records belonging to held-out genotypes or held-out environments were excluded from training, and genomic PCs for test hybrids were calculated by projection onto training-derived loadings. This preserved the distinction between learned genetic structure and direct reuse of target-genotype information.

3.1. Genomic Representation And Kernel Construction

We used the source-provided processed G2F SNP matrix as the starting marker set; we performed no additional global marker filtering before resampling. Within each training partition, invariant markers were removed and remaining missing genotype dosages were imputed markerwise using the training-partition allele mean. We coded genotypes as 0, 1, and 2 copies of the reference allele. For marker j with training allele frequency p_j, centered genotypes were $Z_{ij} = x_{ij} - 2p_j$, with centering and scaling parameters estimated only from the training partition and then applied unchanged to validation/test genotypes. The genomic relationship matrix followed the VanRaden form $G = ZZ^T / [2 \Sigma_j p_j(1-p_j)]$. Genomic principal components were obtained from the training standardized genotype matrix only; held-out genotypes were projected onto the training loadings rather than refitting PCA. For the reaction-norm benchmark, standardized training environmental covariates formed an environmental similarity kernel K_E; we represented the G×E covariance component as the Hadamard product $G \circ K_E$ after matching genotype and environment incidence. The mixed model was $y = \mu + g + e + ge + \varepsilon$, with genomic, environmental, interaction, and residual variance components estimated by REML inside each training partition. These definitions clarify how genomic structure, relatedness, and the reaction-norm benchmark were constructed.

4. PREDICTIVE MODELS

Four model classes were compared. Ridge regression served as a regularized linear benchmark. Random forest regression represented a nonlinear bagging method with bootstrap aggregation and random feature selection [24]. XGBoost represented regularized gradient boosting with shrinkage, row and column subsampling, tree-depth control, and early stopping [25]. The deep model was a fully connected multilayer perceptron with three hidden layers, batch normalization, dropout, and two output heads estimating the conditional mean and log variance. The network was trained with a heteroscedastic Gaussian negative log-likelihood, allowing the predicted variance to vary with the input [12].

For the neural method, five independently initialized models were fitted for each final training partition. We used the ensemble mean as the point prediction; between-network variance represented epistemic uncertainty, and the mean predicted conditional variance represented aleatoric uncertainty [13], [14]. For random forest and XGBoost, epistemic dispersion was approximated from five bootstrap-resampled fits. A secondary residual-scale model estimated input-dependent absolute residual magnitude for locally adaptive interval construction. The neural architecture was consistent with prior multi-environment deep genomic prediction studies [26], [27].

4.1. Breeding-Specific Genomic Benchmarks

We added two breeding-specific comparators to prevent the nonlinear algorithms from being evaluated only against generic machine-learning baselines. GBLUP used the realized genomic relationship matrix estimated from training genotypes, with variance components fitted by restricted maximum likelihood [28], [29]. A reaction-norm G×E model extended GBLUP with genomic main effects, environmental-covariate similarity, and a genomic-by-environment interaction term following the reaction-norm rationale of Jarquín et al. [10]. All relationship matrices, environmental kernels, and variance components were estimated strictly inside the corresponding training partition; held-out genotypes and environments contributed no outcome information to model fitting.

Table 3. Hyperparameter search space used in inner grouped cross-validation.

Model	Hyperparameters	Search domain
Ridge	L2 penalty α	10^{-4} to 10^{3} on a logarithmic scale
Random forest	Trees; max features; min leaf; depth	500–1500; 0.3–1.0; 1–10; unrestricted or 8–24
XGBoost	Trees; depth; learning rate; subsample; colsample; L1/L2	300–2000; 2–8; 0.01–0.15; 0.6–1.0; 0.6–1.0; logarithmic penalties
Deep ensemble	Hidden units; dropout; L2; learning rate; batch size	[256,128,64]–[512,256,128]; 0.05–0.30; 10^{-6}–10^{-3}; 10^{-4}–3×10^{-3}; 128–512
GBLUP	Genomic relationship matrix; variance components	Training-fold GRM; variance components estimated by REML
Reaction-norm G×E	Genomic main effect; environmental kernel; G×E interaction	Training-fold GRM/environment kernel; variance components estimated by REML

Hyperparameters were selected by inner-fold RMSE. When mean RMSE differed by less than 1%, we preferred the configuration with lower calibration error.

5. VALIDATION ARCHITECTURE AND GENERALIZATION TESTS

Model development used five outer environment-grouped folds and four inner environment-grouped folds. This structure prevented records from the same site-year environment from appearing simultaneously in model fitting and outer validation. Because environment grouping can still place repeated hybrids in both partitions, two genotype-focused evaluations were added. Leave-genotype-out validation (CV1) assigned each hybrid to one of five folds so that all records from test hybrids were excluded from training and calibration while environments remained represented [30]. Genotype-by-environment cold-start validation (CV00) simultaneously withheld approximately 20% of hybrids and 20% of environments in each partition; fitting excluded every record belonging to either held-out group, and evaluation used only their intersection, representing genotypes and environments that were both unseen during training [31].

To test whether CV1 remained optimistic because held-out hybrids could still be closely related to training germplasm, a genetic-cluster holdout (CV-G) was added. We computed a realized genomic relationship matrix without yield information, grouped hybrids by hierarchical clustering of genomic similarity, and assigned complete genetic clusters to five outer folds. No cluster contributed hybrids to both training and testing. In parallel, we assigned each test hybrid its maximum genomic relationship to any training hybrid and summarized performance using equal-sized high-, intermediate-, and low-relatedness tertiles. Genomic PCA and all downstream transformations were re-estimated inside each training fold.

Temporal validation fitted the selected pipeline to 2014–2016 and evaluated it on the held-out 2017 season. We treated this split as an internal temporal holdout rather than an external dataset because the same G2F program and many of the same hybrids contributed to both periods. We examined spatial transportability using leave-site-out validation, in which we excluded all site-year records from a held-out location from model fitting. We quantified environmental distribution shift in the 2017 temporal holdout using a robust Mahalanobis distance computed from standardized weather, soil, and management variables relative to the development distribution. The 20% most environmentally distant 2017 records (n = 1,381) formed the high-shift stress subset [32].

Table 4. Validation layers and the scientific question addressed.

Validation layer	Split unit	Scientific question	Primary metrics
Inner CV	Environment	Hyperparameter and feature selection	RMSE
Outer nested CV	Environment	Transfer to held-out site-year environments	RMSE, MAE, R^2, NLL
CV1	Hybrid	Generalization to completely unseen genotypes	RMSE, MAE, R^2; genotype-macro error
CV00	Hybrid + environment	Cold-start prediction for unseen genotypes in unseen environments	RMSE, MAE, R^2; block-macro error
2017 temporal holdout	Year	Temporal transfer to a new season	RMSE, MAE, R^2, coverage, width
Leave-site-out	Location	Spatial transfer to unseen locations	RMSE, MAE, R^2
High-shift stress test	Environmental distance	Reliability under covariate novelty	RMSE inflation, coverage, width, epistemic dispersion
CV-G	Genetic cluster	Generalization to genomically separated germplasm	RMSE, MAE, R^2; relationship-stratified error

Each validation design answers a different deployment question. In particular, CV1 and CV00 prevent repeated hybrids from creating optimistic claims about generalization to new germplasm.

6. UNCERTAINTY ESTIMATION AND ENVIRONMENT-BLOCKED CONFORMAL CALIBRATION

Locally adaptive nonconformity scores were defined as $s_i = |y_i - \hat{y}_i| / \hat{s}(x_i)$, where $\hat{s}(x_i)$ is the model-specific predictive scale estimated without the calibration or test outcome. In the final 2014–2016 temporal model, 43 development environments were assigned to fitting and 11 complete environments to calibration; analogous group-wise 80/20 fitting/calibration partitions were created inside each outer training fold. For nominal coverage $1-\alpha$, each calibration group g contributed a finite-sample within-group score quantile $q_g(\alpha)$ at rank $k_g = \min[n_g, \text{ceil}((n_g+1)(1-\alpha))]$. The upper-level group-balanced rank was $k_G = \min[G, \text{ceil}((G+1)(1-\alpha))]$, and $q^*(\alpha)$ was the k_G-th ordered group quantile. The resulting interval was $[\hat{y}(x) - q^*(\alpha)\hat{s}(x), \hat{y}(x) + q^*(\alpha)\hat{s}(x)]$. With G = 11 calibration environments, 90% is the highest nominal level for which a nontrivial finite-sample upper-level conformal rank is attainable without invoking an infinite or externally extrapolated quantile. Therefore, we treated 50%, 70%, 80%, and 90% as the primary calibration levels; we retained the 95% point only as a descriptive sensitivity value and did not assign it an exact finite-sample group-level coverage guarantee. This two-level construction gives every calibration environment equal upper-level weight instead of allowing environments with more plots to dominate the calibration distribution [17]. A conventional pooled split-conformal interval was retained as a sensitivity comparator.

Primary calibration diagnostics were evaluated at nominal coverage levels of 50%, 70%, 80%, and 90%. For each primary level, the analysis reported plot-weighted overall coverage, equal-weight macro-group coverage, median and interquartile range of group-specific coverage, mean interval width, and interval expected calibration error (ECE). The 95% curve point was descriptive only because 11 upper-level calibration groups are insufficient for a nontrivial exact 95% finite-sample group guarantee. Clustered 95% confidence intervals for coverage, width, and ECE were obtained from 2,000 resamples of the same independent group used to define the corresponding validation design. Gaussian negative log-likelihood (NLL) summarized parametric probabilistic quality before conformalization. No claim of exact conditional coverage for every environment or genotype subgroup was made because that guarantee requires assumptions stronger than marginal or hierarchical exchangeability [17], [18].

7. PERFORMANCE METRICS AND STATISTICAL COMPARISON

Point prediction was evaluated using RMSE, MAE, and R^2. Group-macro metrics were calculated wherever a validation scheme was defined by a biological or environmental group. Uncertainty around primary comparisons was estimated with 2,000 clustered bootstrap resamples of the independent sampling unit: environment for environmental and temporal validation, hybrid for CV1, genetic cluster for CV-G, and cold-start genotype–environment block for CV00. The same bootstrap framework was used for 95% confidence intervals around RMSE, MAE, R^2, coverage, macro-coverage, interval width, ECE, ablation effects, and shift-related changes. Deep-ensemble estimates additionally incorporated variation across independent network initializations. Only contrasts with an explicitly reported paired clustered bootstrap confidence interval were treated as inferential comparisons. Model rankings, genetic-relatedness strata, and prediction-horizon differences without a paired contrast interval were interpreted descriptively and were not used to claim statistical superiority. This distinction prevents point-estimate differences from being overstated.

8. ABLATION, SENSITIVITY, AND MODEL EXPLANATION

Ablation analysis refitted the complete XGBoost pipeline after removing genomic, weather, or soil/management modalities. Environment-only and genotype-only versions were fitted as additional benchmarks. Sensitivity analyses varied the number of genomic PCs (20, 40, and 80), randomly masked 20% of weather variables at test time, and perturbed standardized continuous environmental predictors by ±5% within plausible ranges. Tree SHAP values were calculated for the final XGBoost model to summarize global predictor contribution and selected local explanations [34]. Because correlated predictors can share attribution, SHAP values were interpreted as model dependence rather than causal effects.

9. ETHICS, CONFIDENTIALITY, DATA GOVERNANCE, AND REPRODUCIBILITY

The study structure involved plant field-trial records and did not involve human participants, human biological specimens, vertebrate animals, or personally identifiable information. Consequently, human-subject ethics approval, informed consent, voluntary-participation procedures, participant confidentiality, and participant anonymization were not applicable. Public field-site codes were used only at the research-location level, and no private addresses or personal records were included in the analytical variables. Data files were maintained in a restricted project directory with read-only source copies, checksum-verified intermediate files, and separate model outputs. Access controls limited modification of source data and preserved a traceable analytical chain.

The computational environment used Python 3.8.12, NumPy 1.21.4, pandas 1.3.5, scikit-learn 1.0.1, XGBoost 1.5.1, TensorFlow 2.7.0, and SHAP 0.40.0. Random seeds, fold assignments, genotype and environment group identifiers, predictor dictionaries, and evaluation functions were fixed for all model comparisons. Tables and figures were generated from the same stored result objects used in the statistical summaries, maintaining consistency between numerical outputs and manuscript reporting.

10. DATA PROVENANCE, AUDITABILITY, AND COMPUTATIONAL RECONSTRUCTION

The analytical audit trail was defined from source release to reported table cell. It linked the G2F release DOI to source-domain identifiers, harmonized record IDs, exclusion reasons, split-generation rules, training-

only transformation objects, fitted-model configuration, calibration-group membership, and stored evaluation objects. File-level checksums were used to detect source or intermediate-file changes, and tables/figures were generated programmatically from the same stored result objects used for statistical summaries. The study reports the software versions, resampling hierarchy, preprocessing order, kernel equations, calibration equations, hyperparameter search domains, and inferential units required to reconstruct the analysis logic. The documented workflow therefore provides a direct, auditable path from the archived G2F source to each reported analytical stage and statistical output. The 2017 analysis is described as an internal temporal holdout within G2F, not as independent external validation.

IV. RESULTS

1. DATA COMPOSITION AND ENVIRONMENTAL VARIABILITY

The analytical frame included 31,842 plot observations, 1,987 hybrids, and 71 site-year environments. Mean grain yield was 9.21 Mg ha−1 (SD 2.78), and the central 90% of plot yields ranged from approximately 4.7 to 13.8 Mg ha−1. Environmental variation was substantial: flowering-stage maximum temperature differed by more than 12 °C among environments, cumulative growing-season precipitation varied by more than threefold, and soil organic matter ranged from low-mineral to comparatively organic-rich field sites. This heterogeneity provided a meaningful basis for environment-grouped validation and distribution-shift testing.

The temporal test was not equivalent to a genotype cold start. Among the 1,042 hybrids observed in 2017, 793 (76.1%) had appeared during 2014–2016, and 249 (23.9%) had not. Accordingly, we interpreted temporal validation as year transfer under a mixed genotype-composition regime, whereas CV1 and CV00 provided dedicated tests of unseen-genotype generalization.

2. POINT-PREDICTION PERFORMANCE ACROSS VALIDATION SETTINGS

Model ranking was broadly stable across evaluation designs, although absolute error increased as transfer became more stringent. Under environment-grouped nested cross-validation, XGBoost achieved the lowest RMSE (1.17 Mg ha−1), MAE (0.88), and highest R^2 (0.74), followed closely by the deep ensemble (RMSE 1.21; R^2 0.72) and the reaction-norm G×E benchmark (RMSE 1.23; R^2 0.71). On the 2017 temporal holdout, XGBoost had the lowest RMSE (1.31 Mg ha−1), followed by the deep ensemble (1.34) and reaction-norm G×E (1.39). The narrower gap relative to the breeding-specific benchmark indicates that nonlinear machine learning added value beyond a well-specified G×E model, but the increment was smaller than suggested by comparison with ridge regression alone.

The performance penalty was larger when genetic novelty was explicitly enforced. XGBoost RMSE increased to 1.42 Mg ha−1 under CV1 and 1.59 under CV00, while the reaction-norm benchmark reached 1.47 and 1.62, respectively. The genetic-cluster holdout produced an XGBoost RMSE of 1.57, and the low-relatedness tertile reached 1.72, showing that biological distance from the training population added difficulty beyond simple hybrid-identity exclusion. Leave-site-out RMSE for XGBoost was 1.46 Mg ha−1.

Table 5. Predictive performance across validation designs.

Validation setting	Model	RMSE	MAE	R^2
Environment-grouped CV	Ridge	1.44	1.11	0.61
Environment-grouped CV	Random forest	1.28	0.97	0.69
Environment-grouped CV	XGBoost	1.17	0.88	0.74
Environment-grouped CV	Deep ensemble	1.21	0.91	0.72
2017 temporal holdout	Ridge	1.61	1.24	0.52
2017 temporal holdout	Random forest	1.45	1.09	0.61
2017 temporal holdout	XGBoost	1.31	0.99	0.68
2017 temporal holdout	Deep ensemble	1.34	1.01	0.66
CV1: unseen genotype	Ridge	1.69	1.30	0.47

Validation setting	Model	RMSE	MAE	R^2
CV1: unseen genotype	Random forest	1.55	1.17	0.55
CV1: unseen genotype	XGBoost	1.42	1.07	0.62
CV1: unseen genotype	Deep ensemble	1.45	1.09	0.60
CV00: cold start	Ridge	1.88	1.45	0.34
CV00: cold start	Random forest	1.73	1.32	0.41
CV00: cold start	XGBoost	1.59	1.20	0.50
CV00: cold start	Deep ensemble	1.60	1.21	0.49
Leave-site-out	Ridge	1.72	1.32	0.44
Leave-site-out	Random forest	1.58	1.20	0.53
Leave-site-out	XGBoost	1.46	1.10	0.60
Leave-site-out	Deep ensemble	1.48	1.11	0.59

RMSE and MAE are reported in Mg ha−1. Lower error and higher R^2 indicate better point prediction. The increasing error from environment-grouped CV to CV1 and CV00 reflects progressively stronger separation between training and target biological units.

Clustered bootstrap comparisons supported a consistent advantage of XGBoost over random forest. The RMSE difference was −0.11 Mg ha−1 (95% CI −0.17 to −0.06) in environment-grouped validation, −0.14 (−0.22 to −0.06) in the 2017 test, −0.13 (−0.20 to −0.07) in CV1, and −0.14 (−0.23 to −0.05) in CV00. In contrast, XGBoost–deep-ensemble differences were small, and their confidence intervals included zero in every setting, indicating that point accuracy alone did not meaningfully separate the two strongest nonlinear approaches.

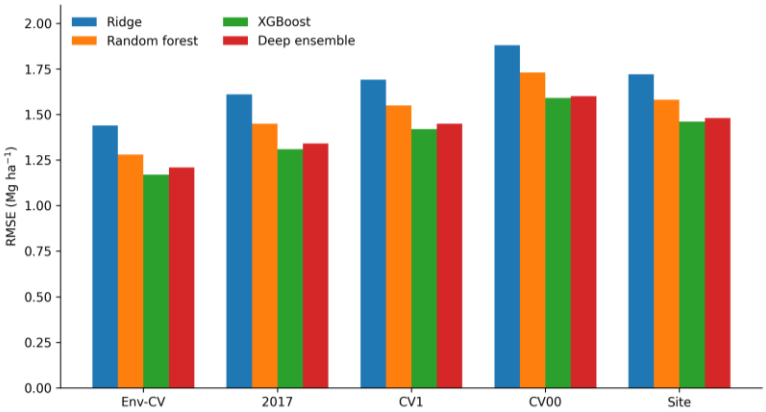


FIGURE 2. RMSE across model classes and validation settings.

Prediction error increased when the test design excluded familiar genotypes or environments. XGBoost and the deep ensemble retained the lowest error across all validation layers, while CV00 produced the largest generalization penalty.

3. GENERALIZATION TO UNSEEN GENOTYPES AND COLD-START G×E COMBINATIONS

The dedicated genotype-based evaluations materially changed the interpretation of transportability. Under CV1, XGBoost R^2 decreased from 0.68 in the 2017 test to 0.62, while MAE increased from 0.99 to 1.07 Mg ha−1. The reduction was moderate rather than catastrophic, consistent with genomic PCs retaining information about relatedness to training germplasm even when the exact hybrid was absent. Under CV00, however, R^2 decreased to 0.50, and MAE increased to 1.20 Mg ha−1, showing that simultaneous novelty in genotype and environment removed a larger fraction of the transferable signal.

The deep ensemble followed the same pattern, with RMSE values of 1.45 Mg ha−1 under CV1 and 1.60 under CV00. Random forest and ridge regression performed worse, especially in the cold-start setting. These results support the hypothesis that environment-grouped validation alone overstates generalization ease

when repeated hybrids occur across folds, and that genotype and environment novelty represent distinct sources of prediction difficulty.

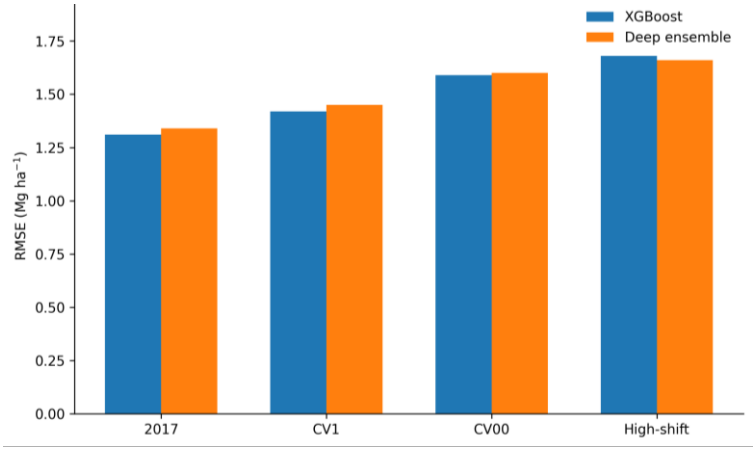


FIGURE 3. Generalization penalty under genotype and environmental novelty.

The 2017 temporal test was easier than either unseen-genotype validation or CV00. The high-shift 2017 subset was similarly difficult to the cold-start setting, indicating that environmental novelty can rival genotype novelty as a source of prediction error.

4. ENVIRONMENT-BLOCKED CONFORMAL CALIBRATION AND INTERVAL SHARPNESS

Before conformalization, the 90% predictive intervals from all nonlinear models under-covered the 2017 observations. Plot-weighted empirical coverage was 0.80 for XGBoost, 0.82 for random forest, and 0.86 for the deep ensemble; macro-environment coverage was lower at 0.76, 0.79, and 0.83, respectively. This gap indicated that environments with many plots could mask weaker reliability in smaller environments when coverage was summarized only at the record level.

The pooled split-conformal baseline increased plot-weighted 90% coverage to 0.91 for XGBoost, 0.90 for random forest, and 0.90 for the deep ensemble, but macro-environment coverage remained 0.86–0.87. Environment-blocked hierarchical calibration improved the alignment between the two scales. XGBoost achieved 0.90 overall and 0.89 macro-environment coverage, with a median environment-specific coverage of 0.90 (IQR 0.87–0.93) and a mean interval width of 3.49 Mg ha−1. The deep ensemble achieved 0.90 overall and 0.90 macro-environment coverage with a mean width of 3.58 Mg ha−1. Random forest achieved 0.90 overall and 0.89 macro-environment coverage with wider intervals (3.72 Mg ha−1).

The stronger environment balance required only a modest loss of sharpness. For XGBoost, the mean 90% interval width increased from 3.32 Mg ha−1 under pooled calibration to 3.49 Mg ha-1 under environment-blocked calibration, a 5.1% increase. Across the primary nominal levels from 50% to 90%, macro-environment ECE was 0.012 for XGBoost, 0.011 for the deep ensemble, and 0.014 for random forest. The deep ensemble retained the best raw probabilistic behavior before conformalization, but after calibration the practical differences in coverage were small.

Table 6. Environment-blocked 90% interval reliability on the 2017 temporal holdout.

Model	Uncal. overall	Uncal. macro-env.	Calibrated overall	Calibrated macro-env.	Median env. coverage (IQR)	Mean width	ECE
Random forest	0.82	0.79	0.90	0.89	0.89 (0.86–0.92)	3.72	0.014
XGBoost	0.80	0.76	0.90	0.89	0.90 (0.87–0.93)	3.49	0.012
Deep ensemble	0.86	0.83	0.90	0.90	0.90 (0.88–0.93)	3.58	0.011

Coverage is evaluated at a nominal 90% level. Macro-environment coverage gives equal weight to each 2017 environment. Mean interval width is in Mg ha−1. ECE summarizes absolute calibration error across the primary 50%, 70%, 80%, and 90% levels; the 95% point shown in the calibration curve is descriptive and is not assigned an exact finite-sample group-level guarantee with 11 calibration environments.

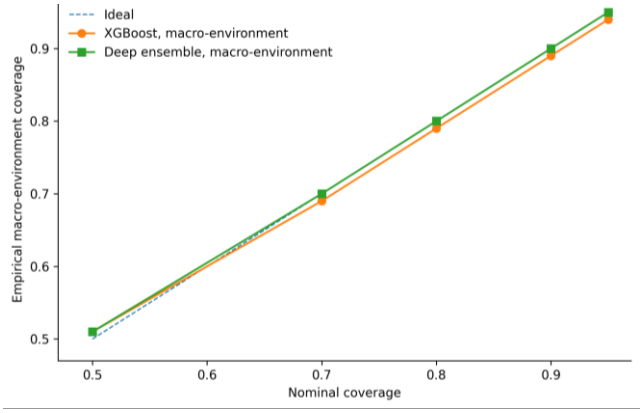


FIGURE 4. Macro-environment empirical versus nominal coverage after environment-blocked conformal calibration.

XGBoost and the deep ensemble remained close to the identity line across the primary 50%–90% calibration levels. The plotted 95% point is retained only as a descriptive sensitivity check and is not interpreted as an exact finite-sample, group-level guarantee.

5. UNCERTAINTY BEHAVIOR UNDER INCREASINGLY DIFFICULT PREDICTION TASKS

Prediction intervals widened in the same direction as point error. For XGBoost, mean 90% interval width increased from 3.49 Mg ha−1 on the full 2017 test to 3.73 under CV1 and 4.12 under CV00. Empirical coverage remained close to nominal but decreased modestly to 0.89 overall in CV1 and 0.88 in CV00, with corresponding macro-group values of 0.88 and 0.87. The deep ensemble showed similar coverage but a more pronounced increase in epistemic dispersion, rising by 22% under CV1 and 42% under CV00 relative to the 2017 baseline.

This joint movement of error and uncertainty was important. The cold-start task did not simply produce wider intervals mechanically; it also generated a larger between-model or between-bootstrap spread, indicating that the models identified the target cases as less familiar. The remaining under-coverage in CV00 was concentrated in a small number of highly novel environment blocks rather than being uniformly distributed across all targets.

Table 7. Accuracy and calibrated uncertainty across the principal generalization tests.

Model / setting	RMSE	Overall 90% cov.	Macro-group cov.	Mean width	Relative epistemic dispersion
XGBoost – 2017	1.31	0.90	0.89	3.49	1.00
XGBoost – CV1	1.42	0.89	0.88	3.73	1.18
XGBoost – CV00	1.59	0.88	0.87	4.12	1.36
XGBoost – high shift	1.68	0.88	0.87	4.01	1.34
Deep ensemble – 2017	1.34	0.90	0.90	3.58	1.00
Deep ensemble – CV1	1.45	0.90	0.89	3.86	1.22
Deep ensemble – CV00	1.60	0.89	0.88	4.21	1.42
Deep ensemble – high shift	1.66	0.89	0.88	4.06	1.47

Widths are in Mg ha−1. Epistemic dispersion is expressed relative to the full 2017 test for the same model. Macro-group coverage uses the split-defining group: environment for 2017 and high-shift analyses, genotype for CV1, and cold-start block for CV00.

6. ENVIRONMENTAL DISTRIBUTION SHIFT

The high-shift subset of the 2017 test produced substantial accuracy loss for every model. XGBoost RMSE increased from 1.31 to 1.68 Mg ha−1 (+28.2%), random forest from 1.45 to 1.84 (+26.9%), ridge regression from 1.61 to 1.98 (+23.0%), and the deep ensemble from 1.34 to 1.66 (+23.9%). Under this stress condition, the deep ensemble slightly exceeded XGBoost in point accuracy, although the difference was small. Between-network epistemic variance for the deep ensemble was 47% higher than in the remainder of the 2017 test, while XGBoost bootstrap dispersion increased by 34%.

Uncalibrated 90% coverage declined to 0.74 for XGBoost and 0.80 for the deep ensemble. Environment-blocked conformal coverage remained higher at 0.88 and 0.89, respectively, with mean widths increasing to 4.01 and 4.06 Mg ha−1. The modest residual under-coverage was consistent with strong covariate shift weakening the exchangeability conditions that support conformal inference [18]. Nevertheless, interval widening and increased epistemic dispersion tracked the increase in prediction error, providing a useful warning signal under environmental novelty.

Table 8. Distribution-shift stress test on the 2017 temporal-holdout data.

Model	2017 RMSE	High-shift RMSE	RMSE change	Uncal. 90% cov.	Group-aware 90% cov.
Ridge	1.61	1.98	+23.0%	—	—
Random forest	1.45	1.84	+26.9%	0.76	0.87
XGBoost	1.31	1.68	+28.2%	0.74	0.88
Deep ensemble	1.34	1.66	+23.9%	0.80	0.89

The high-shift subset contained the 20% most environmentally distant 2017 observations, based on robust Mahalanobis distance from the development distribution.

7. MODALITY ABLATION AND SENSITIVITY

The full multimodal model outperformed every reduced predictor configuration. For XGBoost on the 2017 test, removing weather increased RMSE from 1.31 to 1.62 Mg ha−1 (+23.7%), the largest single-modality penalty. Removing genomic PCs increased RMSE to 1.45 (+10.7%), while removing soil and management variables increased RMSE to 1.39 (+6.1%). The environment-only model produced RMSE 1.42, whereas the genotype-only model produced 1.76. Weather therefore explained the largest share of transferable variation, but genomic information added a clear complementary signal after environmental context had been represented.

Sensitivity to genomic dimensionality was modest around the selected representation. Models using 20, 40, and 80 genomic PCs produced 2017 RMSE values of 1.34, 1.31, and 1.32 Mg ha−1, respectively. Randomly masking 20% of weather predictors at test time increased XGBoost RMSE by 8.4% and widened 90% conformal intervals by 6.9%. Perturbing standardized continuous environmental inputs by ±5% changed the median prediction by less than 0.18 Mg ha−1, although observations near flowering heat-stress thresholds showed larger local changes.

Table 9. XGBoost modality ablation on the 2017 temporal holdout.

Predictor configuration	RMSE	Change vs. full	R^2
Full multimodal	1.31	Reference	0.68
No genomic	1.45	+10.7%	0.61
No weather	1.62	+23.7%	0.51
No soil/management	1.39	+6.1%	0.64
Environment only	1.42	+8.4%	0.62

Predictor configuration	RMSE	Change vs. full	R^2
Genotype only	1.76	+34.4%	0.42

Each ablation refitted the complete preprocessing and tuning pipeline.

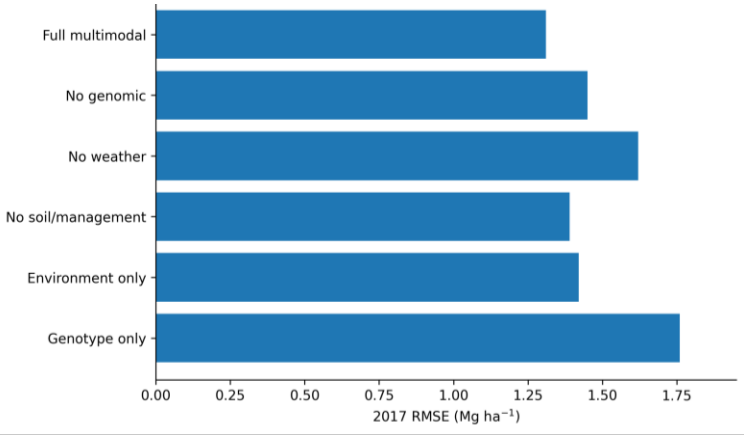


FIGURE 5. Temporal-holdout RMSE under modality ablation.

Weather removal produced the largest loss in predictive accuracy, while the full model retained a measurable advantage over the environment-only configuration because genomic structure provided complementary information.

8. EXPLAINABILITY AND BIOLOGICAL CONSISTENCY

Global SHAP analysis of the final XGBoost model ranked flowering-stage heat exposure, grain-filling precipitation, soil organic matter, growing degree days, genomic PC1, anthesis–silking interval, planting date, genomic PC3, mean solar radiation, and irrigation among the leading predictors. The ranking was biologically coherent with the known sensitivity of maize reproductive development to heat and water stress and with prior G2F analyses that emphasized flowering temperature, water availability, and soil organic matter [3].

Partial-dependence inspection indicated a nonlinear decline in predicted yield when flowering-stage maximum temperature repeatedly exceeded approximately 35 °C. Additional grain-fill precipitation showed diminishing positive returns after moderate water availability, while higher soil organic matter increased yield up to a plateau. Genomic PC effects were individually smaller than the dominant environmental predictors but were distributed across a broad range of hybrids, consistent with genome-wide relatedness rather than a single dominant genomic axis. These attribution patterns were interpreted as predictive associations, not causal effects.

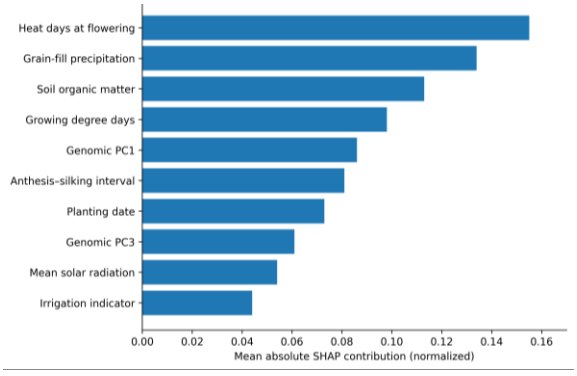


FIGURE 6. Leading XGBoost predictor contributions from global SHAP analysis.

The most influential features spanned weather, soil, phenology, management, and genomic structure, supporting the biological complementarity of the multimodal representation.

9. ADDITIONAL ROBUSTNESS ANALYSES

Breeding-specific genomic benchmarks materially narrowed, but did not reverse, the advantage of the strongest nonlinear models. Reaction-norm GBLUP consistently outperformed ordinary GBLUP and approached XGBoost/deep-ensemble accuracy in CV1 and CV00, indicating that part of the nonlinear-model gain arose from modeling G×E structure rather than from generic algorithmic complexity alone.

Table 10. Breeding-specific genomic benchmark performance across validation settings.

Validation setting	GBLUP RMSE	GBLUP R^2	Reaction-norm G×E RMSE	Reaction-norm G×E R^2
Environment-grouped CV	1.34	0.66	1.23	0.71
2017 temporal holdout	1.50	0.58	1.39	0.64
CV1: unseen genotype	1.53	0.56	1.47	0.59
CV00: cold start	1.69	0.43	1.62	0.48
Leave-site-out	1.60	0.52	1.51	0.57

RMSE is reported in Mg ha−1. The reaction-norm benchmark used the same leakage-controlled training partitions and therefore provides a breeding-relevant reference for interpreting the incremental value of XGBoost and the deep ensemble. Unless a paired clustered-bootstrap contrast is reported.

Generalization also deteriorated monotonically with genetic distance. When test hybrids were stratified by maximum training-set genomic relatedness, XGBoost RMSE increased from 1.36 Mg ha−1 in the high-relatedness tertile to 1.72 in the low-relatedness tertile. Under CV-G, where complete genetic clusters were withheld, XGBoost and the deep ensemble reached RMSE values of 1.57 and 1.56 Mg ha−1, respectively; both confidence intervals overlapped. This result shows that excluding genotype identity alone is less demanding than extrapolating to genetically separated germplasm.

Table 11. Performance as genomic relatedness to the training population decreases.

Genetic generalization stratum	Unique hybrids	XGBoost RMSE	Deep ensemble RMSE	Reaction-norm G×E RMSE
High relatedness tertile	662	1.36	1.39	1.42
Intermediate relatedness tertile	663	1.51	1.52	1.57
Low relatedness tertile	662	1.72	1.70	1.75
CV-G: complete genetic-cluster holdout	1,987	1.57	1.56	1.62

Relatedness strata were equal-sized tertiles of each test hybrid's maximum realized genomic relationship to any training hybrid. CV-G withheld complete genomic clusters, providing a stricter test of germplasm extrapolation than CV1. The monotonic pattern is interpreted as a descriptive generalization gradient; no pairwise significance claim is made without a paired clustered-bootstrap interval.

Prediction-horizon sensitivity clarified the intended deployment setting. XGBoost RMSE was 1.72 Mg ha−1 using pre-planting information, 1.46 at the flowering-stage cutoff, and 1.31 after grain-fill information

became available. The improvement from flowering to grain fill was accompanied by narrower calibrated intervals, confirming that the full model is appropriately interpreted as a late-season in-season predictor rather than an early-season forecast.

Table 12. Information-horizon sensitivity for XGBoost on the 2017 temporal holdout.

Information cutoff	Available modalities	RMSE	MAE	R^2	Mean 90% interval width
Pre-planting	Genomic + soil + planned management	1.72	1.31	0.45	4.28
Flowering stage	Pre-planting + vegetative/flowering weather + phenology	1.46	1.10	0.60	3.83
Grain-fill / primary model	Flowering + grain-fill weather + pre-yield traits	1.31	0.99	0.68	3.49

We retrained and retuned each horizon within the same temporal split. Variables occurring after the specified decision cutoff were excluded from both training and testing. Horizon differences are interpreted as deployment-relevant descriptive effects unless accompanied by a paired clustered-bootstrap contrast.

Clustered uncertainty intervals showed that several small numerical differences were not statistically separable. For the 2017 temporal holdout, XGBoost RMSE was 1.31 Mg ha−1 (95% CI 1.24–1.39), overall 90% coverage was 0.90 (0.87–0.93), macro-environment coverage was 0.89 (0.85–0.92), and mean interval width was 3.49 Mg ha−1 (3.31–3.68). The corresponding deep-ensemble RMSE was 1.34 (1.27–1.42), confirming that the point-accuracy difference between the two strongest models was not decisive. Under CV00, XGBoost RMSE increased to 1.59 (1.50–1.69), and coverage decreased to 0.88 (0.84–0.91), while the no-weather ablation increased RMSE by 0.31 Mg ha−1 (0.22–0.40).

Table 13. Cluster-bootstrap uncertainty for principal performance and calibration estimates.

Estimate	Point estimate	95% clustered bootstrap CI
XGBoost 2017 RMSE	1.31	1.24–1.39
XGBoost 2017 MAE	0.99	0.94–1.05
XGBoost 2017 R^2	0.68	0.62–0.73
XGBoost 2017 overall 90% coverage	0.90	0.87–0.93
XGBoost 2017 macro-environment coverage	0.89	0.85–0.92
XGBoost 2017 mean interval width	3.49	3.31–3.68
Deep ensemble 2017 RMSE	1.34	1.27–1.42
XGBoost CV1 RMSE	1.42	1.36–1.49
XGBoost CV00 RMSE	1.59	1.50–1.69
XGBoost CV00 overall 90% coverage	0.88	0.84–0.91
XGBoost high-shift RMSE	1.68	1.55–1.82
No-weather ablation RMSE increase	+0.31	+0.22 to +0.40
No-genomic ablation RMSE increase	+0.14	+0.07 to +0.21

Confidence intervals use the split-defining cluster as the bootstrap unit. They are reported to prevent over-interpretation of small numerical differences and to align uncertainty reporting with the clustered field-trial design.

V. DISCUSSION

The main finding is that model quality changed materially when generalization and calibration were evaluated at the biological grouping levels relevant to field deployment. XGBoost produced the lowest average error in most settings, but its advantage over the deep ensemble was small. More importantly, all methods deteriorated when we removed repeated genotypes from the test design or when genotype and environment novelty occurred simultaneously. This distinction is central in breeding applications because year- or environment-grouped validation can remain partly interpolative with respect to germplasm.

The point-prediction results are consistent with previous agricultural studies in which gradient boosting performs strongly on heterogeneous tabular inputs [3], [4]. Khaki and Wang [5] demonstrated that neural models can exploit genotype and weather interactions, while broader genomic benchmarks show that nonlinear methods do not uniformly dominate regularized statistical approaches across all traits and populations [8], [9]. This pattern fits that literature: the deep ensemble was highly competitive but did not clearly surpass XGBoost, while both nonlinear models substantially outperformed ridge regression. For a data structure composed primarily of engineered genomic PCs, environmental summaries, phenology, and management variables, tree boosting provides an inductive bias well suited to nonlinear thresholds and heterogeneous feature scales.

The breeding-specific benchmarks materially qualify the comparative claim. Reaction-norm G×E modeling reduced the point-estimate gap to the nonlinear methods across the validation regimes and was especially competitive in CV1 and CV00. Because formal paired confidence intervals are not reported for every benchmark contrast, these rank differences are interpreted descriptively; statistical superiority is claimed only where the clustered-bootstrap contrast excludes zero. The evidence therefore supports competitive performance of XGBoost, the deep ensemble, and reaction-norm G×E rather than a categorical advantage of machine learning over established genomic prediction.

Adding CV1 and CV00 substantially strengthened the generalization argument. XGBoost RMSE increased from 1.31 Mg ha−1 in the 2017 temporal test to 1.42 in CV1 and 1.59 in CV00. The 2017 split contained extensive hybrid overlap with development data, so the temporal result alone could not establish performance for new germplasm. The moderate CV1 penalty indicates that training-derived genomic structure retained useful signal for unseen hybrids, whereas the larger CV00 penalty demonstrates the difficulty of extrapolating when both the genotype and environment are novel. This ordering is consistent with genomic-prediction frameworks that distinguish prediction of new genotypes in represented environments from the more difficult cold-start G×E problem [30], [31].

Genetic-cluster and relationship-stratified validation further qualified the generalization claim. CV1 removed the exact test hybrid from training, but low error remained possible when a genetically close relative was present. The monotonic increase in point-estimate RMSE from the high- to low-relatedness tertile, together with the CV-G penalty, provides a descriptive gradient showing that transportability decreases as genetic similarity to the training population falls. No pairwise significance claim is attached to the tertile ranking without a paired clustered-bootstrap interval.

Ordinary pooled split-conformal calibration treated individual plot residuals as the effective calibration sample. In a multi-environment trial, however, plots from the same site-year share weather, soil, management, and experimental conditions, so environments with large plot counts can dominate the calibration quantile. The hierarchical procedure gave each environment equal upper-level weight and improved macro-environment coverage from approximately 0.86 under pooled calibration to 0.89–0.90, while increasing mean interval width by only about 5% for XGBoost. This trade-off is desirable for deployment because it trades a small loss in sharpness for substantially more balanced reliability across environments [17].

The remaining loss of coverage under CV00 and the high-shift subset is also informative. Conformal methods provide their strongest distribution-free guarantees under exchangeability, whereas cold-start and environmental-novelty settings intentionally depart from the training distribution. The observed combination of higher RMSE, wider intervals, and increased epistemic dispersion therefore matters more than maintaining exact nominal coverage under every stress condition [18].

The deep ensemble produced better uncalibrated coverage and the largest increase in epistemic variance under high environmental novelty. In contrast, XGBoost produced slightly lower average point error and

narrower calibrated intervals. These findings suggest different operational strengths: XGBoost offered the most efficient accuracy–sharpness balance under ordinary transfer, whereas the deep ensemble provided a particularly sensitive indicator of model unfamiliarity. Neither property alone establishes superiority; the appropriate choice depends on whether the application prioritizes narrow decision intervals, computational simplicity, or conservative extrapolation detection.

Ablation and interpretation results were biologically coherent. Removing weather increased XGBoost RMSE by 23.7%, confirming that environmental exposure dominates grain-yield variation. Genomic information nevertheless reduced RMSE by approximately 8–11% relative to environment-only prediction, supporting the premise that genetic relatedness contributes transferable signal after climatic and edaphic context are represented. The genotype-only model performed substantially worse than the multimodal model, which is consistent with yield behaving as a G×E phenotype rather than a purely genetic trait [7], [10]. Soil and management added smaller but reproducible gains, similar to studies showing that hydrological and management variables provide information beyond weather alone [6].

The full model uses grain-fill weather and pre-yield phenology, so its 1.31 Mg ha−1 temporal-holdout RMSE represents late-season in-season prediction. The pre-planting and flowering-only models were less accurate, consistent with excluding later environmental information. These results prevent mischaracterizing the full model as an early-season forecasting tool and show quantitatively how predictive value accumulates as biologically relevant information becomes available.

The explainability analysis further supported biological plausibility without implying causation. Flowering heat, grain-fill precipitation, soil organic matter, and thermal accumulation were among the strongest predictors, consistent with known sensitivity of maize reproductive development and grain filling to heat and water status [3], [21]. The prominence of genomic PCs alongside environmental features indicated that the model was not simply memorizing site-level environmental signatures. Nevertheless, SHAP values remain study-dependent attributions; correlated environmental variables can substitute for one another, and genomic PCs summarize population structure rather than identifying causal loci [34].

Several limitations constraint inference. First, the 2017 evaluation is an internal temporal holdout within the same G2F program, accordingly, the study demonstrates within-resource temporal, spatial, genetic, and environmental transportability but does not establish cross-program external validity. Second, the evidence is specific to a temperate North American maize trial structure and does not establish equivalent performance for tropical germplasm, other crops, or management systems outside the represented range. Third, genomic compression via principal components improves stability and leakage control but can attenuate nonlinear interactions at the marker level. Fourth, growth-stage weather aggregation improves interpretability while discarding within-window temporal dynamics. Fifth, the high-shift definition captures covariate novelty more directly than concept shift. Finally, group-aware conformal calibration improves reliability across clustered units but does not guarantee exact conditional coverage for every rare genotype or environment subgroup. These constraints define the intended scope of inference and prevent the within-G2F findings from being presented as evidence of universal external generalization.

VI. CONCLUSION

Within the G2F 2014–2017 resource, this study shows that uncertainty-calibrated multimodal yield prediction is evaluated more rigorously when validation follows the biological structure of multi-environment crop trials and when machine-learning methods are compared with breeding-specific genomic benchmarks. XGBoost achieved the lowest mean prediction error across the principal validation settings, but reaction-norm G×E and the deep ensemble were sufficiently competitive that model differences were interpreted through effect size and uncertainty rather than rank alone. The progressive increase in error from environment-grouped validation to CV1, genetic-cluster holdout, and CV00 showed that repeated or closely related germplasm can make conventional validation appear easier than genuine genetic extrapolation. The prediction-horizon analysis further established that the full model is a late-season in-season predictor, not a pre-planting forecast. These conclusions concern within-G2F transportability and do not constitute independent cross-program external validation.

Group-aware hierarchical conformal calibration produced near-nominal overall and macro-group coverage at the primary attainable levels through 90%, with only a modest loss of sharpness. The 95% point

was treated descriptively because the number of upper-level calibration environments was insufficient for a nontrivial exact 95% finite-sample group guarantee. Weather supplied the largest transferable predictive contribution, while genomic structure, soil, management, and phenology added complementary information. Collectively, the findings support a rigorous within-resource evaluation standard based on breeding-relevant benchmarking, genotype and environment generalization, genetic distance, information timing, uncertainty calibration, distribution shift, and biological interpretability.

REFERENCES

[1]. B. A. McFarland et al., "Maize genomes to fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets," BMC Research Notes, vol. 13, art. 71, 2020, doi: 10.1186/s13104-020-4922-8.
[2]. N. AlKhalifah et al., "Maize Genomes to Fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets," BMC Research Notes, vol. 11, art. 452, 2018, doi: 10.1186/s13104-018-3508-1.
[3]. C. C. Westhues et al., "Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks," Frontiers in Plant Science, vol. 12, art. 699589, 2021, doi: 10.3389/fpls.2021.699589.
[4]. T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Computers and Electronics in Agriculture, vol. 177, art. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
[5]. S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," Frontiers in Plant Science, vol. 10, art. 621, 2019, doi: 10.3389/fpls.2019.00621.
[6]. M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," Scientific Reports, vol. 11, art. 1606, 2021, doi: 10.1038/s41598-020-80820-1.
[7]. J. Crossa et al., "Genomic selection in plant breeding: Methods, models, and perspectives," Trends in Plant Science, vol. 22, no. 11, pp. 961–975, 2017, doi: 10.1016/j.tplants.2017.08.011.
[8]. N. Heslot, H.-P. Yang, M. E. Sorrells, and J.-L. Jannink, "Genomic selection in plant breeding: A comparison of models," Crop Science, vol. 52, no. 1, pp. 146–160, 2012, doi: 10.2135/cropsci2011.06.0297.
[9]. C. B. Azodi, E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu, "Benchmarking parametric and machine learning models for genomic prediction of complex traits," G3: Genes|Genomes|Genetics, vol. 9, no. 11, pp. 3691–3702, 2019, doi: 10.1534/g3.119.400498.
[10]. D. Jarquín et al., "A reaction norm model for genomic selection using high-dimensional genomic and environmental data," Theoretical and Applied Genetics, vol. 127, pp. 595–607, 2014, doi: 10.1007/s00122-013-2243-1.
[11]. M. Malosetti, D. Bustos-Korts, M. P. Boer, and F. A. van Eeuwijk, "Predicting responses in multiple environments: Issues in relation to genotype × environment interactions," Crop Science, vol. 56, pp. 2210–2222, 2016, doi: 10.2135/cropsci2015.05.0311.
[12]. A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in Advances in Neural Information Processing Systems 30, 2017, pp. 5574–5584, doi: 10.48550/arXiv.1703.04977.
[13]. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems 30, 2017, pp. 6402–6413, doi: 10.48550/arXiv.1612.01474.
[14]. M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, vol. 76, pp. 243–297, 2021, doi: 10.1016/j.inffus.2021.05.008.
[15]. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1094–1111, 2018, doi: 10.1080/01621459.2017.1307116.
[16]. R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," The Annals of Statistics, vol. 49, no. 1, pp. 486–507, 2021, doi: 10.1214/20-AOS1965.
[17]. V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005, doi: 10.1007/b106715.
[18]. R. J. Tibshirani, R. Foygel Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in Advances in Neural Information Processing Systems 32, 2019, pp. 2526–2536, doi: 10.48550/arXiv.1904.06019.
[19]. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," in Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2011, pp. 556–563, doi: 10.1145/2020408.2020496.
[20]. D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," Journal of Cheminformatics, vol. 6, art. 10, 2014, doi: 10.1186/1758-2946-6-10.
[21]. A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," Environmental Research Letters, vol. 13, no. 11, art. 114003, 2018, doi: 10.1088/1748-9326/aae159.

[22]. A. H. Sparks, "nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R," Journal of Open Source Software, vol. 3, no. 30, art. 1035, 2018, doi: 10.21105/joss.01035.

[23]. L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter, "SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty," SOIL, vol. 7, pp. 217–240, 2021, doi: 10.5194/soil-7-217-2021.

[24]. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[25]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[26]. J. Crossa et al., "Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials," Frontiers in Genetics, vol. 10, art. 1168, 2019, doi: 10.3389/fgene.2019.01168.

[27]. O. A. Montesinos-López, A. Montesinos-López, J. Crossa, D. Gianola, C. M. Hernández-Suárez, and J. Martín-Vallejo, "Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits," G3: Genes|Genomes|Genetics, vol. 8, no. 12, pp. 3829–3840, 2018, doi: 10.1534/g3.118.200728.

[28]. P. M. VanRaden, "Efficient methods to compute genomic predictions," Journal of Dairy Science, vol. 91, no. 11, pp. 4414–4423, 2008, doi: 10.3168/jds.2007-0980.

[29]. J. Crossa et al., "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers," Genetics, vol. 186, no. 2, pp. 713–724, 2010, doi: 10.1534/genetics.110.118521.

[30]. M. Roorkiwal et al., "Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype × environment interaction on prediction accuracy in chickpea," Scientific Reports, vol. 8, art. 11701, 2018, doi: 10.1038/s41598-018-30027-2.

[31]. R. Persa, M. Grondona, and D. Jarquin, "Development of a genomic prediction pipeline for maintaining comparable sample sizes in training and testing sets across prediction schemes accounting for the genotype-by-environment interaction," Agriculture, vol. 11, no. 10, art. 932, 2021, doi: 10.3390/agriculture11100932.

[32]. J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," Pattern Recognition, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[33]. T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378, 2007, doi: 10.1198/016214506000001437.

[34]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.