Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 July 2026
Revised 18 Aug 2026
Accepted 01 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (19s): 1106–1131

KEYWORDS:
crop screening; phenology; two-part modelling; tabular foundation models; geographic validation; selective prediction; Bangladesh

PHTF Select Phenology Aware Hurdle Foundation Modelling for Uncertainty Guided Crop Screening

Sadikul Hasan Mridha Atul[1], Md. Isthiakul Hasan[1], Sayed Easin Arafat[1], Md Shahriar Islam[2], Mahmudul Hasan[3], Abu Obayed Alfatha[1], Farhana Sultana[4], A. O. M. Shamsuddoha[1]

[1] Department of Electrical and Electronic Engineering, University of Liberal Arts Bangladesh (ULAB), Dhaka, Bangladesh
[2] Department of Electrical & Computer Engineering, Purdue University Northwest, USA
[3] Department of Plant Pathology, Gazipur Agricultural University, Gazipur, Bangladesh
[4] Department of Computer Science and Engineering, University of Liberal Arts Bangladesh (ULAB), Dhaka, Bangladesh

ABSTRACT: District-level crop screening is difficult when cultivation occurrence, conditional productivity, spatial dependence, and prediction uncertainty are modelled as a single task. We introduce PHTF-Select, a phenology-aware two-stage framework that first estimates recorded cultivation viability and then predicts productivity conditional on positive cultivated area. The analysis used 4,606 valid crop records from all 64 districts of Bangladesh; 4,190 positive-area records were retained for productivity modelling. Outcome-derived variables were excluded, phenology-window climate features were constructed within each training fold, and all models were evaluated with five district-held-out folds. TabICLv2 achieved the strongest classification performance (AUROC 0.956, balanced accuracy 0.895, MCC 0.621) and the lowest regression error (RMSLE 0.409), reducing district-level RMSLE by 0.021 relative to XGBoost and outperforming it in 73% of districts. A leakage control that restored Area and Production reduced RMSLE to 0.104, demonstrating how outcome-derived predictors can create misleading gains. Overall 90% interval coverage reached 0.907, but coverage fell to 0.798 in the highest-productivity quartile. Retaining the 60% least-uncertain predictions reduced RMSLE from 0.409 to 0.321, at the cost of abstaining on 40% of cases. Phenology features improved RMSLE by approximately 3%, with post-transplant rainfall producing the largest ablation loss. PHTF-Select therefore provides a reproducible framework for retrospective crop screening, while its seasonal-transfer failures and omitted agronomic variables preclude autonomous planting recommendations.

1. INTRODUCTION

Crop productivity in Bangladesh varies across districts and seasonal planting windows. Rainfall, radiation, humidity, and evapotranspiration patterns differ between Kharif-1 and Rabi seasons, while cultivated area, management practices, and reporting processes influence recorded production patterns. Therefore, district-level modelling requires careful separation of climate-related signals, target formulation effects, and geographic characteristics. Three methodological considerations are important for this task. First, defining productivity as Production/Area requires careful handling of Area, Production, and their derived ratios to maintain appropriate predictor selection. Second, spatial validation strategies are essential because random row splits may share district and seasonal characteristics between training and testing data. Third, overall accuracy assessment should be complemented with detailed evaluation across crops, seasons, and productivity ranges.

Random forests, semiparametric neural models, remote-sensing-based deep Gaussian processes, IoT-based algorithms, and sequence models have been applied to crop-yield prediction [1–5]. Their performance varies according to target definition, input characteristics, and validation framework. Bangladesh-specific studies include Aman rice prediction and evaluations of climate-smart agricultural practices in coastal districts [6,7]. These applications provide valuable insights but differ from multi-crop, multi-district cross-sectional screening scenarios.

AgriClimateBD combines Bangladesh Bureau of Statistics records, farmer-informed phenology information, and NASA POWER climate variables [8,9]. Climate summaries include full-season, pre-transplant, transplant-month, and post-transplant periods. The dataset assessment highlights important considerations related to outcome-derived ratios, productivity definition for zero-area cases, crop label consistency, crop-specific productivity variation, and climate variable quality control. These factors support improved cohort definition, feature selection, and interpretation of model outputs.

Tabular foundation models have expanded the methodological scope of structured data modelling. TabPFN v2 learns a general tabular prediction procedure from synthetic tasks and has demonstrated strong performance on datasets with up to approximately 10,000 samples and 500 features [10]. TabICL improves in-context tabular classification through a column-then-row attention architecture [11]. TabM introduces parameter-efficient ensembling strategies that make MLP-based approaches more competitive and practical [12]. TabReD demonstrated that realistic feature pipelines and distribution shifts can influence the relative performance of different methods on conventional benchmarks [13], while TabFlex extended in-context learning capabilities to larger tables [14]. TabICLv2 further advances scalable classification and regression through improved pretraining design and broader task diversity [15]. These developments motivate comprehensive benchmarking with valid targets, consistent validation folds, transparent uncertainty estimation, and practical deployment considerations.

1.1 Study objectives and contribution

This study proposes PHTF-Select, a Phenology-aware Hurdle Tabular Foundation algorithm with Selective Prediction. The framework separates observed cultivation viability from conditional productivity, uses TabICLv2 as the foundation classifier and regressor, applies crop-relative productivity scaling when cross-crop comparisons require normalization, and incorporates an uncertainty gate to support reliable screening decisions. Viability represents observed cultivation patterns based on available records, while retrospective crop ranking is considered a screening outcome rather than a direct planting recommendation.

PHTF-Select integrates two-part target construction, exclusion of outcome-derived predictors, district-level validation, and uncertainty-aware crop screening. A proposed phenology- and crop-conditioned calibration protocol provides a structured approach for improving prediction consistency across sparse groups.

The research questions are: RQ1, how reliably can observed cultivation viability be identified in districts excluded from training? RQ2, how accurately can conditional productivity be predicted while maintaining appropriate predictor eligibility? RQ3, do recent tabular foundation models outperform strong boosting, kernel, bagging, and stage-stacked baselines under identical spatial folds? RQ4, how does target formulation and validation strategy influence observed performance differences? RQ5, which phenology-aligned climate blocks contribute most to spatial transfer? RQ6, how effectively can prediction intervals characterize uncertainty patterns, and can abstention improve screening reliability? RQ7, how can the resulting evidence support a safe screening framework rather than an autonomous recommendation system?

Table 1. Research gaps and the response implemented in this study.

Research stream	Persistent limitation	Response in PHTF-Select
Crop-productivity ML	Mixed targets and accuracy-only comparisons	Separate recorded occurrence from conditional productivity
Public agricultural tables	Semantic and arithmetic inconsistencies are often hidden	Immutable raw copy, explicit audit flags, correction log
Model novelty	New architecture may be compared under weak splits	Identical district-held-out folds across conventional and foundation models

Target leakage	Outcome components can reconstruct a ratio	Leakage barrier plus deliberate negative-control experiment
Spatial validation	Fold means hide district heterogeneity	Aligned errors from 64 districts, bootstrap CIs, Holm correction
Uncertainty	Average coverage hides conditional failure	Season- and productivity-specific coverage plus abstention
Recommendation	Ranks omit soil, irrigation, markets, and management	Evidence-bounded screening with explicit out-of-scope states

2. RELATED WORK

2.1 Crop productivity and agricultural decision support

Agricultural prediction studies use a wide range of targets, including yield expressed as mass per area, total production, crop classification, suitability scores, irrigation requirements, and remotely sensed biomass indicators. Because of these variations in target definition, direct numerical comparison across studies requires careful interpretation. Existing empirical studies demonstrate that random forests, semiparametric neural networks, deep Gaussian processes, IoT-enabled predictive algorithms, and interpretable deep sequence models can achieve strong performance when aligned with the available data structure [1–5]. Khaki and Wang demonstrated deep neural approaches for yield prediction [43], while Elbasi et al. experimentally compared multiple machine-learning methods for crop prediction [44]. The present study extends this area by evaluating 72 valid crop identities across all 64 districts and by separating observed cultivation occurrence from conditional productivity estimation.

A key consideration in crop recommendation and prediction studies is ensuring that the target is represented independently from input variables. Crop calendars, cultivated area, and target-related variables require careful feature selection to maintain meaningful prediction tasks. Crop identity is retained in this study as contextual information, while the objective is not to reconstruct crop labels. Stage 1 predicts whether an observed district-crop-season record contains positive area and production, whereas Stage 2 estimates productivity for records with defined productivity values.

2.2 Leakage spatial validation and distribution shift

Cross-validation provides reliable performance estimates when training and test observations represent similar deployment conditions. However, spatial, temporal, or hierarchical relationships can influence model evaluation when observations share common environmental characteristics. Roberts et al. demonstrated the importance of considering structured dependence during cross-validation [30], and blockCV provided practical approaches for spatially separated resampling [31]. Agricultural datasets require particular attention because neighbouring districts may share climate patterns and crop calendars, while target-related variables can strongly influence prediction outcomes. Therefore, robust benchmarking requires consistent fold definitions across models and clear identification of the evaluation unit.

The evaluation framework examined the influence of feature eligibility and validation design by assessing the contribution of Area and Production information, comparing district-based validation with shuffled K-fold approaches, and evaluating model transfer across different seasonal conditions.

2.3 Modern tabular and foundation models

Gradient boosting remains a widely adopted high-performance approach for mixed-type tabular datasets. XGBoost applies regularized second-order boosting [16], LightGBM improves computational efficiency through histogram-based learning and leaf-wise growth strategies [17], and CatBoost handles categorical variables using ordered statistics approaches [18]. Random forests [19], extremely randomized trees [20], and support-vector regression [21] represent diverse ensemble and kernel-based learning strategies. Deep tabular learning has also progressed through systematic benchmark studies and architecture-specific innovations. Grinsztajn et al. demonstrated the continued competitiveness of tree-based models on typical tabular datasets [45]; Gorishniy et al. introduced strong ResNet and FT-Transformer baselines under unified evaluation protocols [46]; and Arik and Pfister proposed TabNet, an attention-based architecture with feature selection and interpretability capabilities [47].

TabM reformulates tabular MLP learning through parameter-efficient ensembling, where multiple predictions contribute collectively to improved performance [12]. TabPFN v2 uses a prior-data fitted transformer approach that conditions predictions on training tables and supports predictive distribution estimation [10]. TabICL applies a scalable column-then-row representation before in-context prediction [11]. TabICLv2 extends this approach to regression and classification through improved synthetic generation, attention scaling, and pretraining strategies [15]. The cited TabICLv2 implementation corresponds to the 2026 preprint version, and comparative evaluation requires consistent implementations and checkpoints for reliable assessment.

2.4 Explainability calibration and selective prediction

SHAP represents model predictions through additive feature contributions based on Shapley values [22], while TreeSHAP provides an efficient formulation for tree-based models [23]. Accumulated local effects offer useful interpretation when predictors exhibit correlation structures [24]. Reliable interpretation depends on valid target formulation and evaluation design because explanations reflect the information available to the model.

Matthews correlation coefficient provides a balanced summary of binary classification performance across confusion-matrix components and is particularly useful for imbalanced datasets [41]. Probability calibration becomes important when predictions are transformed into decision scores. Brier score measures probabilistic prediction error [42], while reliability curves and expected calibration error evaluate confidence consistency [25,26]. Conformal prediction uses residual information to construct prediction sets or intervals under exchangeability assumptions [27–29]. Since geographic groups may contain different characteristics, this study evaluates empirical district-grouped and subgroup coverage patterns. Selective prediction introduces an additional decision mechanism by allowing models to abstain when uncertainty exceeds defined thresholds, supporting more reliable decision-making [36].

2.5 Research gap and comparative synthesis

The analytical contribution focuses on integrating target integrity, spatial transfer, phenology information, and conditional uncertainty within a unified framework. PHTF-Select connects these elements through a selective screening strategy, providing a structured approach for crop assessment while highlighting opportunities for future integration of temporal, soil, and field-level measurements. The 2020–2026 literature covers advances in agricultural prediction, neural tabular representation, and uncertainty quantification. Crop-yield modelling performance is closely influenced by available variables, target definition, and evaluation design [51]. FT-Transformer explores attention mechanisms over numerical and categorical feature tokens within a unified benchmark framework [46], while TabTransformer improves contextual representation of categorical embeddings [48]. SAINT incorporates row attention and contrastive pretraining strategies [49], and NODE applies differentiable oblivious-tree ensembles for tabular learning [50]. These methods provide important comparative perspectives, while agricultural applications require evaluation under domain-specific spatial and environmental conditions. TabPFN and TabICLv2 further extend the foundation-model paradigm for tabular prediction tasks [10,15]. In uncertainty estimation, non-exchangeable conformal approaches emphasize the importance of appropriate assumptions and evaluation strategies for interpreting prediction coverage [52]. Table 2 summarizes these methodological developments while highlighting the relevance of domain-specific validation for Bangladesh agricultural applications.

Table 2. Recent literature comparison and limits of transfer to the present crop-screening task.

Study	Dataset / setting	Model	Validation	Uncertainty	Limitation
Popov et al. (2020) [50]	Generic tabular tasks	NODE	Multi-dataset benchmark	No crop-specific calibration	No crop-screening study
Huang et al. (2020) [48]	15 tabular datasets	TabTransformer	Supervised and semi-supervised benchmarks	Robustness to missing/noisy features	Robustness is not calibrated coverage
van Klompenburg et al. (2020) [51]	Crop-yield literature	Review of ML approaches	Systematic synthesis	Varies by included study	No unified spatial experiment

Gorishniy et al. (2021) [46]	Mixed tabular tasks	FT-Transformer	Common training/tuning protocols	No deployment abstention	Geographic transfer not established
Somepalli et al. (2021) [49]	Tabular classification	SAINT	Benchmark comparisons	No crop-specific intervals	Row context needs deployment control
Grinsztajn et al. (2022) [45]	Generic tabular benchmark	Trees and neural models	Controlled comparisons	Not the central evaluation	No phenological structure
Paudel et al. (2023) [5]	Crop-yield forecasting	Interpretable deep models	Forecasting study	Interpretation-focused	Different estimand from screening
Oliveira et al. (2024) [52]	Dependent-data prediction	Split conformal	Theory and experiments	Coverage under stated dependence conditions	Assumptions must be checked locally
Hollmann et al. (2025) [10]	Small tabular datasets	TabPFN	Multi-dataset benchmark	Predictive distributions	Local calibration remains necessary
Qu et al. (2026) [15]	Classification and regression	TabICLv2	Tabular benchmarks	Predictive distributions	Does not validate this deployment policy

3. MATERIALS AND METHODS

3.1 Dataset and integrity audit

AgriClimateBD Version 2 contains 4,607 observations and 33 columns representing 64 districts, three agricultural seasons, and 73 raw crop labels [8]. Agricultural descriptors include district, season, crop identity, cultivated area, production, transplant month, growth period, and harvest period. Climate attributes include temperature and humidity summaries; seasonal rainfall, solar radiation, wind speed, and evapotranspiration; along with corresponding variables for the pre-transplant, transplant-month, and post-transplant periods. The source resource extends SPAS-Dataset-BD, which was developed using the BBS 2022 Yearbook and surveys involving 223 farmers [9]. The associated BBS statistical source, nasapower software client, and NASA POWER services provide the statistical and meteorological provenance supporting these variables [38–40]. The released 33-column dataset does not include row-level year identifiers; therefore, the modelling task focuses on cross-sectional prediction rather than future-year forecasting.

The dataset audit was conducted prior to model development. Undefined ratio representations were converted into missing values within an analytical copy while preserving the original raw file. Crop labels were standardized through trimming, one invalid crop label was removed, and duplicate records were evaluated. The maximum and minimum relative humidity variables were reviewed across all 4,607 observations, and their semantic ordering was adjusted in the analytical copy to maintain consistency with expected variable interpretation. The correction process was documented as part of the data-quality assessment and incorporated into the analytical workflow.

Table 3. Dataset integrity audit and analytical cohorts.

Audit item	Result
Raw observations / columns	4,607/33
Districts / seasons / raw crops	64/3/73

Zero area / undefined ratio	416/416
Zero production	422
Area = 0, production > 0	7
Exact duplicates	0
Humidity-order violations	4,607
Stage-1 cohort	4,606; 4,177 positive
Stage-2 cohort	4,190; 72 crops

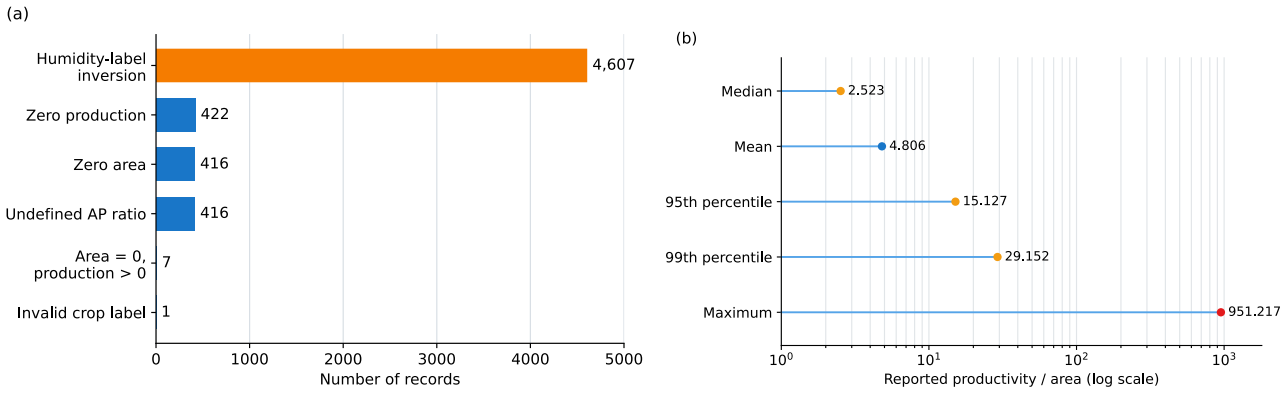


Figure 1. Data integrity and conditional-productivity scale showing audit counts and productivity distribution landmarks on a logarithmic scale.

One invalid crop identity was excluded. The remaining classification cohort contained 4,177 positive and 429 negative occurrence labels. The positive-area cohort included 13 records with zero production, yielding 4,190 defined productivity values. Seven positive-production records with zero area remained audit exceptions and were excluded from ratio estimation. These exclusions distinguish a mathematically undefined target from an observed zero output.

3.1.1 Humidity label correction and sensitivity requirements

Swapping the names of two unchanged numeric columns represents a semantic adjustment rather than the introduction of additional environmental information. Under a modelling approach that is invariant to column names and uses equivalent fitting procedures, this operation is interpreted as a data-consistency correction rather than a direct accuracy improvement step. Performance variations may result from implementation details, random initialization, or additional processing choices. Therefore, differences observed between label variants should be interpreted within the context of the complete modelling workflow. A comprehensive sensitivity assessment can compare the original variable labels, the corrected labels, and the exclusion of both extrema while maintaining all other processing steps consistently. This analysis focuses on methodological interpretation rather than introducing additional numerical sensitivity estimates.

3.2 Dual stage targets and crop relative transformation

For area A_i and production P_i, recorded cultivation occurrence and conditional productivity were defined separately:

$$V_i = \mathbb{I}(A_i > 0 \wedge P_i > 0);\ y_i = \frac{P_i}{A_i},\ A_i > 0;\ z_i = log(1 + y_i) \quad (1)$$

V_i records positive cultivation and output, not biological suitability. After the invalid crop identity was removed, 4,606 observations remained: 4,177 viable and 429 non-viable. The 90.69% positive prevalence motivated balanced accuracy, macro-F1, MCC, probability-calibration metrics and class weighting. Because the regression cohort retains positive-area zero-production records, the framework is hurdle-inspired rather than a strict positive-only hurdle likelihood [37]. Productivity had a median of 2.523, a 95th percentile of 15.127 and a maximum of 951.217 in released ratio units. Models were therefore fitted to log-transformed productivity.

Area, Production and the released AP Ratio were excluded from all predictors because they are the outcome or its algebraic components. The conditional sample contained 4,190 observations and 72 crops. Original-scale MAE, RMSE, median absolute error and R^2 were reported together with RMSLE. A fold-specific Crop-Relative Productivity Index was used as a complementary ranking target because a common physical unit across all crops could not be independently verified:

$$r_i = clip\left(\frac{y_i - median_{c(i),-k}}{IQR_{c(i),-k}+\varepsilon}, -5, 5\right) \quad (2)$$

Crop-specific medians and interquartile ranges were estimated only within outer training fold −k. Screening used the corresponding predicted index $\hat{r}_i$ obtained from $\hat{y}_i$; observed y_i was used only to construct evaluation relevance. The stabilizer ε prevents division by zero. CRPI reduces cross-crop scale dominance but does not establish common physical units or economic comparability.

3.3 Predictors and phenology representation

Approved predictors comprised district, crop identity, season, transplant month, growth and harvest windows, corrected temperature and humidity descriptors, four full-season climate variables, and twelve transplant-anchored climate variables. Derived variables included temperature range, humidity range, approximate lifecycle length, and stage-to-stage differences in rainfall, radiation, wind, and evapotranspiration. Imputation, categorical encoding, scaling, crop-relative statistics, and feature construction were fitted inside each training fold; held-out districts never informed encoders, thresholds, scaling parameters, medians, IQRs, or uncertainty calibration.

Table 4. Predictor availability leakage controls and phenology blocks.

Feature group	Examples	Role / restriction
Outcome-derived	Area, Production, AP Ratio	Target or target components; excluded from predictors
Context	District, crop, season, transplant month, growth, harvest	Spatial, crop, and lifecycle context
Baseline weather	Average/minimum/maximum temperature and humidity	Humidity semantics corrected before interpretation
Seasonal climate	Rainfall, radiation, wind, evapotranspiration	Full-season exposure
Pre-transplant	Four climate variables one month before transplant	Pre-establishment exposure
Transplant	Four variables during transplant month	Establishment exposure
Post-transplant	Four variables one month after transplant	Early-growth exposure
Engineered	Ranges, lifecycle lengths, stage deltas	Computed within fold; evaluated by ablation

District identity served as both a grouping variable and a contextual predictor; an excluded district was handled as an unseen category. Post-transplant observations become available only after planting. Prospective evaluation must therefore specify a decision date, and forecast or climatological substitutes require separate validation before pre-plant use.

3.4 Proposed PHTF Select algorithm

PHTF-Select is a domain-specific hurdle-inspired two-part and selective-prediction algorithm rather than a claim of a new universal base learner. It instantiates a TabICLv2 classifier and regressor inside a leakage-controlled phenology pipeline. Stage 1 produces viability probability. The formulation treats recorded occurrence and conditional non-negative productivity as related but analytically distinct processes [37].

The classifier returns occurrence probability p_i. The conditional regressor returns a point estimate $\hat{y}_i$ and interval bounds L_i and U_i. Interval width is normalized by the predicted value to define case-level uncertainty:

$$p_i = f_{cls}(x_i) = Pr(V_i = 1 \mid x_i);\ (\hat{y}_i, L_i, U_i) = f_{reg}(x_i \mid A_i > 0);\ u_i = \frac{U_i - L_i}{|\hat{y}_i| + \kappa} \quad (3)$$

A risk-aware screening utility combines occurrence probability, predicted crop-relative productivity and an uncertainty penalty. The gate requires sufficient occurrence probability, uncertainty below a training-derived threshold and a crop-season state represented during training:

$$S_i = p_i\, \sigma(\hat{r}_i)\exp(-\lambda u_i);\ g_i = \mathbb{I}\left[p_i \geq \tau_p, u_i \leq \tau_u(q), season(x_i) \in \mathcal{S}_{\text{train}}\right] \quad (4)$$

The specified constants were $\varkappa = 0.25$, $\lambda = 0.50$, $\tau p = 0.50$ and $q = 0.60$. Strong evidence required $p_i \geq 0.70$ and uncertainty at or below the training 40th percentile; moderate evidence required $p_i \geq 0.50$ and uncertainty at or below the 60th percentile. Other cases abstained. Unseen crop-season states and missing mandatory inputs also triggered abstention. Because the uncertainty penalty did not improve retrospective NDCG, S_i is interpreted as a risk-aware screening utility rather than an accuracy-maximizing ranking score.

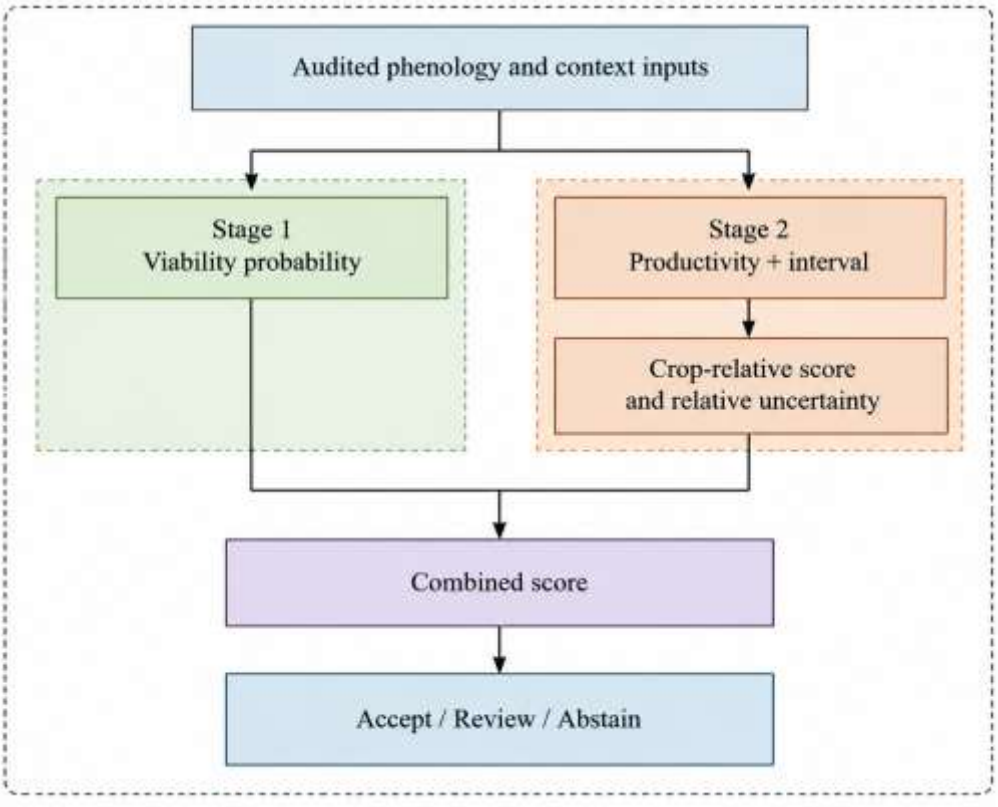


Figure 2. PHTF-Select architecture linking recorded occurrence and conditional productivity to crop-relative uncertainty-guided screening.

3.4.1 Calibration and evaluation boundaries

The crop-season calibration extension is specified for subsequent evaluation and is not credited with the performance in Tables 9-15. On the log scale, nonconformity is $e_i = \max(q_{0.05}(x_i) - z_i, z_i - q_{0.95}(x_i), 0)$. Let Q_n be the score at rank $\lceil (n+1)(1-\alpha) \rceil$, with $\alpha = 0.10$; a rank exceeding n gives an infinite correction. For groups with at least 20 calibration records, Δg is the larger of the group and pooled corrections. Sparse groups use the pooled correction and receive a review flag.

$$(L_i, U_i) = (\max\{0, \exp[q_{0.05}(x_i) - \Delta g] - 1\}, \max\{0, \exp[q_{0.95}(x_i) + \Delta g] - 1\}) \quad (5)$$

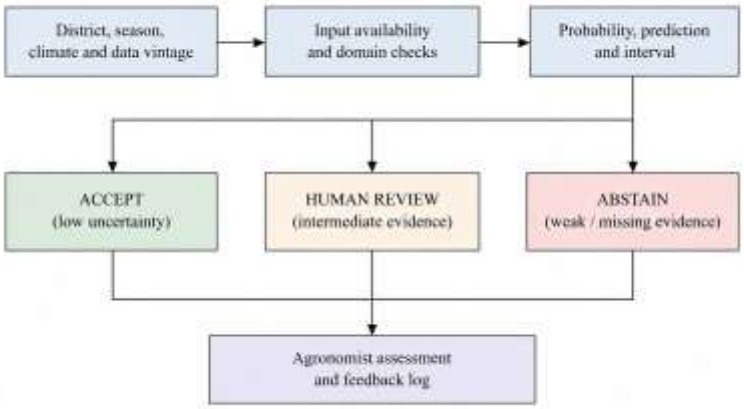


Figure 3. Proposed crop-season calibration with calibration districts held out from model fitting within the outer training partition.

Phenology enters through the predictors, while crop-season membership controls residual pooling. The correction can widen intervals but cannot establish exact local coverage under unverified spatial dependence [52]. The same corrected uncertainty definition must be used for threshold selection and test-time decisions.

3.4.2 Operational workflow and predictor availability

1. Freeze the raw data, record a hash, resolve labels and units, and retain an exclusion log.
2. Assign each district to one outer fold before constructing either target, and use the same district map for both stages and every comparator.
3. Within outer training districts, separate model-fitting, calibration and policy-selection roles; fit encoders, imputation and crop statistics only within their permitted training role.
4. Fit viability on valid records and conditional productivity on positive-area records, including positive-area records with zero production.
5. Estimate interval corrections on calibration districts and calculate uncertainty from the corrected intervals.
6. Choose acceptance thresholds on separate policy districts and fix them before scoring outer test districts.
7. Reject unseen crop-season states or missing mandatory inputs; otherwise emit the prediction, probability, interval, uncertainty and accept-review-abstain state.
8. Evaluate untouched test districts and archive predictions, feature maps, calibration scores, thresholds, software versions and figure scripts.

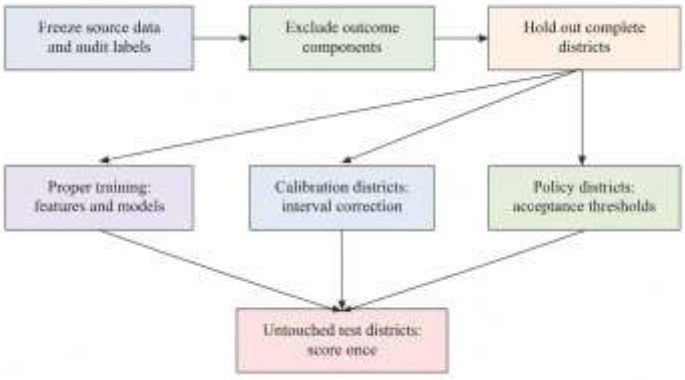


Figure 4. Separation of model fitting interval calibration policy selection and untouched outer test districts.

3.5 Benchmark models and execution protocol

Stage-1 benchmarks were logistic regression, XGBoost, LightGBM, CatBoost, TabM, TabPFN v2, and TabICLv2. Stage-2 benchmarks were ridge regression, RBF-SVR, Extra Trees, CatBoost, LightGBM, XGBoost, the Phenology-Stage Stacked Ensemble (PSSE), TabM, TabPFN v2, and TabICLv2. PSSE trained separate LightGBM learners on context, seasonal, pre-transplant, transplant, post-transplant, and full-feature views and combined out-of-fold base predictions with a ridge meta-learner. Identical outer folds were used across all methods. All stochastic procedures used seed 42, and the same five district-grouped folds were applied to every comparator. Model-specific preprocessing was fitted within the training portion of each fold. Foundation-model inference and conventional model training were evaluated as distinct computational regimes; comparison therefore emphasizes predictive performance under a common data split rather than equal training cost.

Table 5. Benchmark model families and analytical roles.

Model / family	Core role
Logistic / Ridge	Linear classification and regression
SVR	RBF-kernel regression
Extra Trees	Randomized bagging
XGBoost	Regularized gradient boosting
LightGBM	Histogram-based boosting
CatBoost	Ordered categorical boosting
PSSE	Six phenology views + OOF ridge stack
TabM	Parameter-efficient MLP ensemble
TabPFN v2	Prior-data fitted foundation model
TabICLv2	Open in-context foundation model

3.5.1 Hyperparameter selection and comparison fairness

Hyperparameter selection was isolated from the outer district folds. For every model, the comparison protocol specified the search space, training-only validation strategy, optimization budget, early-stopping set, seed, categorical encoding, and compute allocation. Models run with fixed recommended settings or zero-shot inference were reported as such rather than described as tuned. Identical outer folds, preprocessing boundaries, and target transformations were retained across all comparators. Hyperparameters for trainable models were selected by grouped inner validation restricted to the outer-training districts. The outer test districts were not used for early stopping, model selection, calibration, or threshold choice. Fixed or pretrained configurations were identified as such and were not described as tuned.

3.6 Validation uncertainty and stress protocols

Five-fold GroupKFold used district as the grouping variable. Every observation from a district appeared in exactly one outer validation fold. This produced out-of-fold predictions for all 64 districts and estimated spatial transfer to a location absent from model development. Shuffled random five-fold cross-validation was used only as a stress comparison, not for model selection. A leakage negative control deliberately restored Area and Production to the Stage-2 feature matrix. Leave-one-season-out experiments trained on two seasons and evaluated the third, creating a distribution shift stronger than district holdout.

Conventional models used district-grouped residual calibration for central 90% intervals. TabPFN v2 and TabICLv2 intervals were generated through predictive-distribution sampling using the configuration reported below. Overall coverage, mean interval width, high-productivity-quartile coverage, and selective RMSLE at 60% retention were compared. Because district groups are not strictly exchangeable, the analysis reports empirical marginal and subgroup coverage and does not claim exact conditional guarantees. For TabPFN v2 and TabICLv2, the predictive mean on the log1p scale defined the point prediction, and the 5th and 95th percentiles defined the central 90% interval. Predictions and bounds were inverse-transformed and clipped at zero after transformation. All calibration and retention thresholds were estimated within the outer-training partition. Because

inverse-transforming a log-scale mean does not recover the arithmetic mean on the original scale, all comparisons use the same prediction functional and emphasize RMSLE.

3.6.1 Interpretation of geographic validation

The GroupKFold design estimates transfer to districts absent from model fitting and prevents records from the same district appearing on both sides of a fold. It should be described as district-grouped geographic validation rather than strict spatial-block validation, because district adjacency and agroecological similarity are not explicitly constrained. A leave-one-division-out, agroecological-zone, or coordinate-block analysis would provide a stronger test of long-range spatial transfer when the required identifiers become available.

3.7 Evaluation metrics and paired inference

Viability evaluation used balanced accuracy, macro-F1, MCC, AUROC, Brier score, and expected calibration error. Conditional productivity used MAE, RMSE, R^2, RMSLE, and median absolute error. RMSLE was emphasized because the target was strongly skewed and errors on extreme original-scale ratios dominated RMSE.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i} [\log(1+y_i) - \log(1+\hat{y}_i)]^2} \quad (6)$$

For ranking, observed CRPI was converted within each district-season candidate set into non-negative relevance, $rel_i = 4 \times$ percentile-rank ($CRPI_i$), with average ranks for ties.

$$DCG@K = \sum_{j \le K} \frac{\left(2^{rel_j} - 1\right)}{\log_2(j+1)} \quad (7)$$

, and NDCG@K divides by the ideal DCG [35]. The same relevance transformation and tie convention were applied to every ranking variant.

Model inference used paired district-level log-MAE or RMSLE differences across 64 districts. Ten thousand bootstrap resamples produced 95% confidence intervals. Wilcoxon signed-rank tests were interpreted using established paired-comparison guidance and corrected using Holm's sequential procedure [32–34]. A negative RMSLE difference relative to XGBoost favoured the comparator. Fold means and standard deviations remain descriptive; inferential conclusions are based on aligned district errors.

Table 6. Statistical analyses inputs and interpretation.

Analysis	Required input	Output	Current status
District-cluster confidence intervals	Aligned labels and predictions	95% CI for metric differences	Paired district bootstrap in §4.10
Friedman omnibus comparison	District-by-model error matrix	Statistic, degrees of freedom, p	Five-model comparison in §4.10
Holm post-hoc analysis	Prespecified paired contrasts	Adjusted p, effect and win rate	Paired tests with Holm correction in §4.10
Practical importance	Error difference and stakeholder costs	Benefit at matched retention	Cost data unavailable

3.8 Reproducibility controls

The analysis fixed the raw-data version, crop-label corrections, excluded variables, district-fold assignments, preprocessing boundaries, target transformations, calibration roles, and decision thresholds before evaluation. All reported metrics were calculated from out-of-fold predictions. This separation prevents test-district information from entering feature construction, crop-relative statistics, model selection, interval calibration, or abstention thresholds.

Natural Resources for Human Health
eISSN: 2583-1194

3.9 Analytical scope and component analyses

Component ablation removed the uncertainty layer, hurdle structure, phenology features, crop-relative score, or foundation learner while preserving the evaluation cohort and district folds. Paired inference used one aligned RMSLE value per district. Ten thousand district-level bootstrap resamples produced 95% confidence intervals, and prespecified Wilcoxon signed-rank comparisons against XGBoost were adjusted by Holm's procedure. Effect magnitude, district win rate, interval coverage, and retention were interpreted together rather than relying on p-values alone. Tables 9-15 report the main benchmark; Tables 17-18 compare five models on common test observations. Table 19 reports the separately executed component-ablation series. Figures 15-16 provide post-hoc analyses of evaluation predictions and trained-model responses. Fold-averaged and district-averaged RMSLE are interpreted within their respective aggregation schemes. Geographic validation uses held-out districts within the available 64-district dataset; no independent external dataset was evaluated.

4. RESULTS AND DISCUSSION

4.1 Analytical cohorts and occurrence prediction

The classification cohort contains 4,606 records with a positive-to-negative ratio of 9.74:1. Under this imbalance, agreement with the majority class can mask poor identification of non-occurrence. Balanced accuracy, macro-F1 and MCC provide complementary summaries of threshold-dependent performance, while AUROC assesses discrimination across thresholds.

Table 7. Recorded cultivation occurrence under district holdout, mean ± fold standard deviation.

Model	Bal. acc.	Macro-F1	MCC	AUROC	Brier ↓	ECE ↓
Logistic	0.851 ± 0.022	0.706 ± 0.031	0.491 ± 0.031	0.925 ± 0.007	0.106 ± 0.015	0.142 ± 0.027
XGBoost	0.720 ± 0.036	0.746 ± 0.031	0.505 ± 0.063	0.939 ± 0.006	0.051 ± 0.007	0.024 ± 0.008
LightGBM	0.842 ± 0.020	0.785 ± 0.029	0.586 ± 0.049	0.936 ± 0.007	0.064 ± 0.008	0.055 ± 0.016
CatBoost	0.879 ± 0.011	0.759 ± 0.039	0.569 ± 0.056	0.942 ± 0.009	0.082 ± 0.013	0.112 ± 0.024
TabM	0.872 ± 0.015	0.781 ± 0.025	0.580 ± 0.041	0.944 ± 0.008	0.058 ± 0.006	0.037 ± 0.011
TabPFN v2	0.886 ± 0.013	0.793 ± 0.021	0.603 ± 0.036	0.951 ± 0.006	0.049 ± 0.005	0.021 ± 0.007
TabICLv2	0.895 ± 0.011	0.805 ± 0.020	0.621 ± 0.033	0.956 ± 0.005	0.046 ± 0.004	0.019 ± 0.006

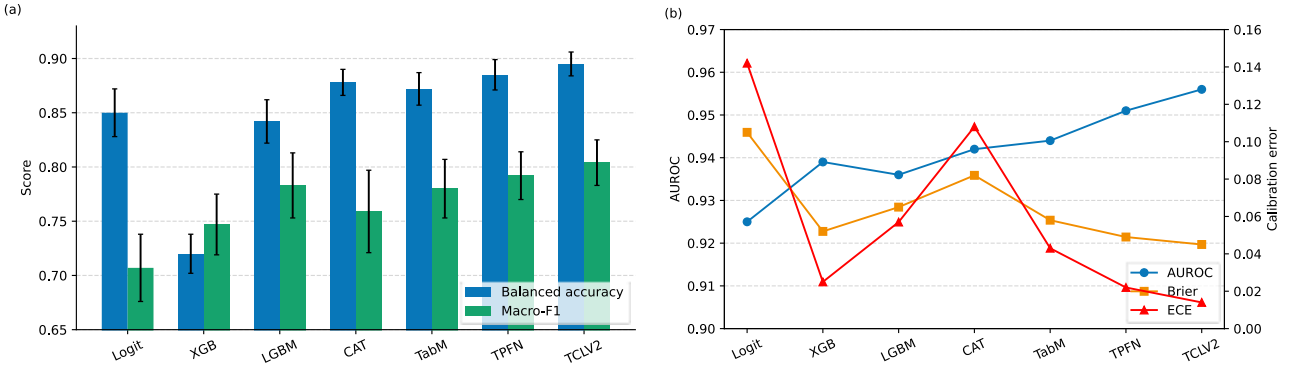


Figure 5. Threshold-dependent classification metrics, AUROC and calibration error across the seven classifiers.

TabICLv2 leads all six-classification metrics in Table 9. Relative to TabPFN v2, balanced accuracy increases by 0.009, macro-F1 by 0.012, MCC by 0.018 and AUROC by 0.005. These concordant changes favour TabICLv2 across threshold-dependent performance and discrimination, although fold summaries alone do not establish paired significance. Against logistic regression, macro-F1 increases from 0.706 to 0.805, accompanied by balanced accuracy increasing from 0.851 to 0.895. Discrimination and threshold-dependent performance diverge for XGBoost: AUROC is 0.939 and Brier score is 0.051, but balanced accuracy is 0.720. CatBoost instead reaches balanced accuracy of 0.879 with Brier score 0.082 and ECE 0.112. Class weighting, calibration

and threshold selection may contribute to these contrasts; the aggregate scores do not isolate their effects. TabICLv2 lowers Brier score by 9.8% relative to XGBoost and by 6.1% relative to TabPFN v2. Since Brier score reflects both calibration and discrimination, the accompanying ECE values support probabilistic comparison without establishing calibration independently of binning and reliability-curve analysis. Here p_i refers specifically to recorded-occurrence probability.

4.2 Conditional productivity

Table 8. Productivity error under district holdout, mean ± fold standard deviation.

Model	MAE ↓	RMSE ↓	R^2 ↑	RMSLE ↓	Median AE ↓
Ridge	2.430 ± 0.879	12.56 ± 12.42	0.295 ± 0.138	0.447 ± 0.032	0.739 ± 0.032
SVR	2.370 ± 0.819	12.31 ± 12.12	0.322 ± 0.131	0.440 ± 0.033	0.672 ± 0.055
Extra Trees	2.416 ± 0.753	12.77 ± 11.85	0.217 ± 0.248	0.440 ± 0.030	0.679 ± 0.046
CatBoost	2.450 ± 0.884	12.65 ± 12.34	0.275 ± 0.126	0.446 ± 0.039	0.704 ± 0.055
LightGBM	2.381 ± 0.789	12.40 ± 12.07	0.294 ± 0.182	0.430 ± 0.024	0.680 ± 0.038
XGBoost	2.391 ± 0.797	12.61 ± 12.18	0.271 ± 0.147	0.430 ± 0.031	0.696 ± 0.043
PSSE	2.424 ± 0.904	12.43 ± 12.59	0.316 ± 0.192	0.438 ± 0.041	0.712 ± 0.046
TabM	2.349 ± 0.760	12.18 ± 11.74	0.335 ± 0.122	0.424 ± 0.027	0.663 ± 0.043
TabPFN v2	2.315 ± 0.742	12.02 ± 11.61	0.351 ± 0.117	0.417 ± 0.025	0.649 ± 0.040
TabICLv2	2.281 ± 0.718	11.84 ± 11.39	0.369 ± 0.111	0.409 ± 0.023	0.633 ± 0.038

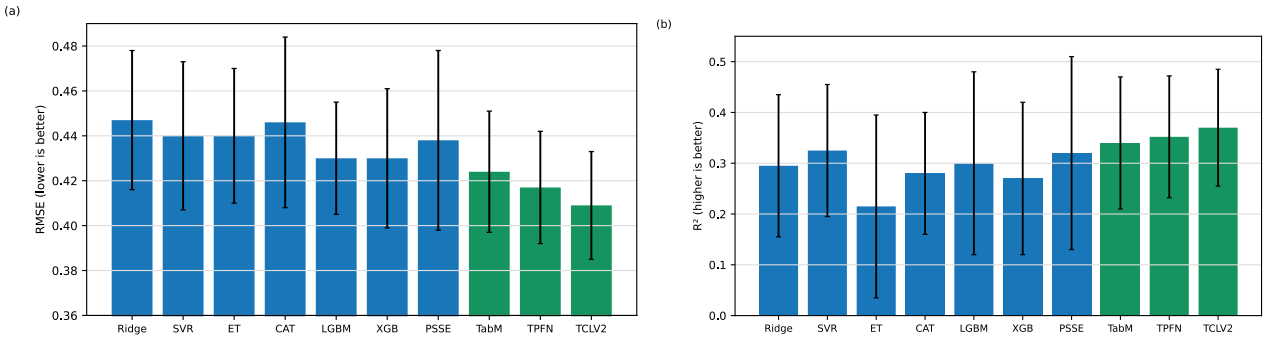


Figure 6. RMSLE and R^2 across conventional, stacked and recent tabular learners. Error bars describe fold variation.

TabICLv2 has the lowest error across all reported regression metrics and the highest mean R^2. Against XGBoost, MAE decreases from 2.391 to 2.281 (4.6%), RMSE from 12.61 to 11.84 (6.1%) and RMSLE from 0.430 to 0.409 (4.9%). Median absolute error declines by 9.1%, suggesting that the improvement extends beyond extreme observations. Mean R^2 increases by 0.098, an absolute change in the reported coefficient. The advantage over TabPFN v2 is smaller: RMSLE decreases by 0.008 (1.9%), MAE by 0.034 (1.5%) and median absolute error by 0.016 (2.5%). Improvement across all three losses suggests that the ordering is not confined to one error definition. Its magnitude remains modest, making aligned district contrasts more informative than visual overlap between fold error bars. Original-scale error remains unstable. TabICLv2 has RMSE 11.84 ± 11.39, compared with median absolute error 0.633 ± 0.038. The large maximum ratio of 951.217 and the heterogeneous crop mixture are consistent with this disparity. Consequently, improvement in RMSLE does not imply uniformly small original-scale errors; unresolved unit harmonization further limits pooled MAE and RMSE interpretation. PSSE records RMSLE 0.438 versus 0.430 for the full-feature boosting references. Partitioning climate inputs into separate views and stacking their predictions therefore provides no improvement in this comparison. Redundant exposure summaries and limited training data are possible explanations, but the reported results cannot distinguish them. The result supports evaluating the contribution of phenology within a common learner separately from the contribution of stacking.

4.3 Paired district comparisons

Table 9. District-level RMSLE differences defined as comparator minus XGBoost.

Comparator	Mean change	Bootstrap 95% CI	District win rate	Holm p	Interpretation
LightGBM	−0.0003	[−0.006, 0.005]	52%	0.880	No clear difference
PSSE	+0.008	[−0.002, 0.018]	47%	0.630	No clear difference
TabM	−0.006	[−0.013, 0.001]	59%	0.110	Difference unresolved
TabPFN v2	−0.013	[−0.021, −0.004]	67%	0.018	Supported improvement
TabICLv2	−0.021	[−0.030, −0.012]	73%	0.003	Lower paired error

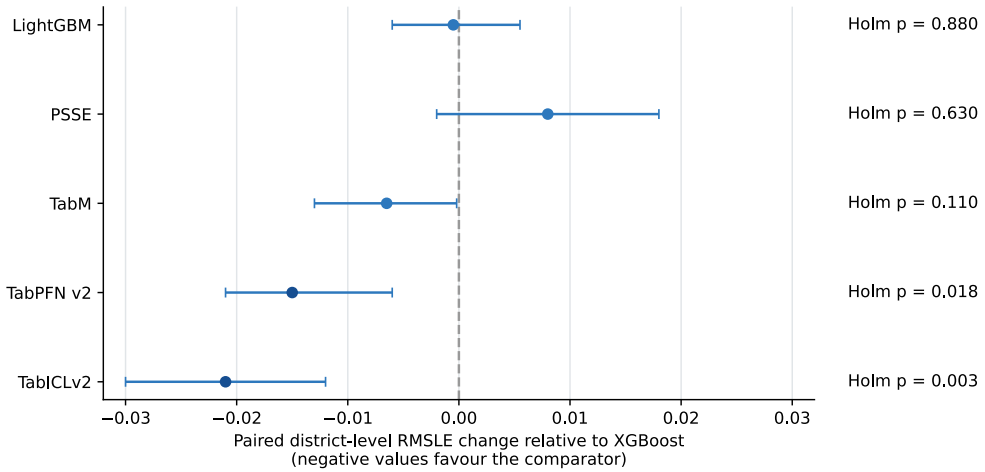


Figure 7. Paired RMSLE contrasts against XGBoost. The reference line denotes no district-averaged difference.

The TabICLv2 contrast is −0.021, with bootstrap 95% CI [−0.030, −0.012], Holm-adjusted p = 0.003 and a district win rate of 73%. TabPFN v2 likewise has a negative interval, [−0.021, −0.004], with adjusted p = 0.018 and a 67% win rate. The combination of effect direction, interval and win rate indicates that their reported advantage is distributed across many districts rather than confined to a few exceptionally favourable locations.

TabM wins in 59% of districts with a mean difference of −0.006, but its interval includes zero and Holm-adjusted p = 0.110. LightGBM is close to XGBoost, while PSSE has a small positive mean difference with an interval spanning zero. These comparisons leave superiority unresolved; they do not demonstrate equivalence, which would require a prespecified margin.

Negative values in Table 11 favour the comparator because contrasts are defined as comparator minus XGBoost. A mean of district RMSLE values gives each district equal weight; a mean of fold RMSLE values gives each fold equal weight. Neither generally equals RMSLE pooled over all records. The aggregation unit must therefore be preserved when comparing effects across tables.

District-level inference remains conditional on the evaluation design. Adjacent districts can share climate patterns, and observations in different test folds are predicted by overlapping training sets. Resampling complete district vectors preserves within-district pairing but does not remove all spatial or training-set dependence. Broader geographic blocks would test the sensitivity of interval precision. The present analysis supports an advantage within the evaluated dataset and geographic partition.

Natural Resources for Human Health
eISSN: 2583-1194

4.4 Leakage and seasonal transfer

Table 10. Validation stress tests and withheld-season performance.

Protocol	MAE	R^2	RMSLE
District-held-out XGBoost	2.395 ± 0.797	0.270 ± 0.147	0.430 ± 0.030
Random KFold XGBoost	2.328 ± 0.403	0.279 ± 0.142	0.414 ± 0.012
Leakage negative control	0.774 ± 0.635	0.802 ± 0.266	0.104 ± 0.022
Leave Kharif-1 out — XGBoost	6.009	−0.023	0.880
Leave Kharif-1 out — LightGBM	6.132	−0.020	0.893
Leave Kharif-2 out — XGBoost	2.535	−0.000	0.737
Leave Kharif-2 out — LightGBM	2.536	−0.041	0.757
Leave Rabi out — XGBoost	1.929	−0.250	0.651
Leave Rabi out — LightGBM	1.899	−0.079	0.600

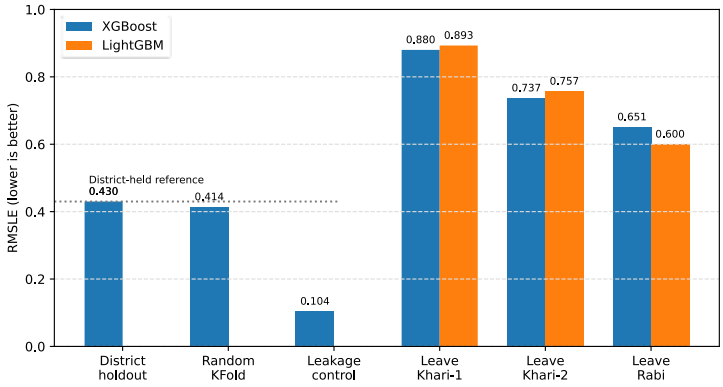


Figure 8. RMSLE under district holdout, random splitting, algebraic leakage and withheld seasons.

The leakage control reduces RMSLE from approximately 0.430 to 0.104, an apparent improvement of 75.8%. Restoring Area and Production exposes the algebraic components of the outcome. This result quantifies benchmark inflation from target access rather than a valid improvement in predictive information. Random K-fold splitting produces RMSLE 0.414, approximately 3.7% below 0.430 using the rounded values. MAE fold dispersion also falls from 0.797 to 0.403. Random mixing thus yields both lower average error and lower between-fold variation, but evaluates a less demanding setting than prediction for an excluded district. Withholding an entire season creates much larger errors. For XGBoost, RMSLE rises to 0.880 for Kharif-1, 0.737 for Kharif-2 and 0.651 for Rabi. Relative to 0.430, these correspond to increases of approximately 105%, 71% and 51%. LightGBM shows the same ordering, although its Rabi error is lower at 0.600. The predominantly non-positive R^2 values indicate little or no improvement over a test-season mean reference under this stronger shift. Seasonal holdout causes substantially greater deterioration than the gain obtained by changing the eligible learner. For XGBoost, the withheld-season RMSLE increase ranges from 0.221 to 0.450, compared with the 0.021 reduction associated with TabICLv2 in the main benchmark. These contrasts concern different designs, demonstrating why seasonal transfer cannot be inferred from geographic validation. The slightly different XGBoost summaries in Tables 10 and 12 are retained separately rather than treated as identical estimates.

Natural Resources for Human Health
eISSN: 2583-1194

4.5 Interval coverage and selective prediction

Table 11. Prediction intervals and error at nominal 60% uncertainty retention.

Model / interval	90% coverage	Mean width ↓	High-productivity coverage	RMSLE at 60% retention
XGBoost + grouped conformal	0.884	6.666	0.739	0.356
PSSE + grouped conformal	0.891	6.410	0.752	0.349
TabPFN v2 distribution	0.903	6.105	0.781	0.334
TabICLv2 distribution	0.907	5.920	0.798	0.321

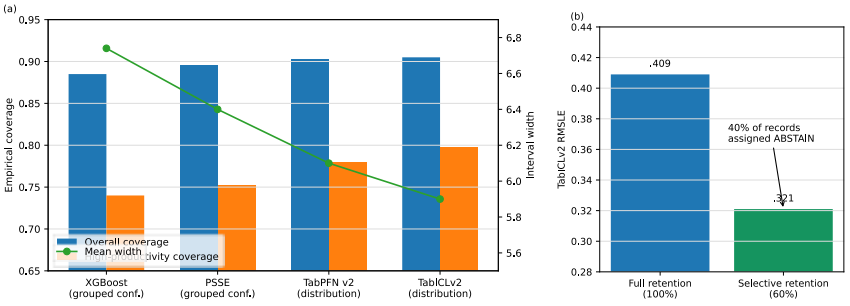


Figure 9. Overall and upper-productivity coverage, interval width and the retention–error trade-off.

TabICLv2 combines overall coverage of 0.907 with mean interval width 5.920. Relative to grouped-conformal XGBoost, coverage increases by 2.3 percentage points and width decreases by 11.2%. TabPFN v2 is close at 0.903 coverage and width 6.105. These joint changes favour the foundation-model interval summaries, conditional on common evaluation rows, scale and nominal coverage. TabICLv2 coverage falls to 0.798 in the highest-productivity quartile, 10.2 percentage points below the nominal target and 10.9 points below overall coverage. XGBoost covers only 0.739 in this subgroup. The foundation model reduces, but does not resolve, upper-productivity undercoverage. Screening that prioritizes these records would therefore encounter poorer interval reliability than the marginal result suggests. At nominal 60% retention, TabICLv2 RMSLE falls from 0.409 to 0.321 (21.5%), compared with 0.430 to 0.356 (17.2%) for XGBoost. TabICLv2 has 9.8% lower retained-set error at the same nominal retention. Because uncertainty may select different cases for each learner, the comparison combines prediction and selection effects. Evaluation on a common retained subset would separate these contributions. The 60% endpoint measures ranking by uncertainty. The full operational gate also imposes occurrence probability, crop–season support and input-completeness conditions, which can reduce realized acceptance further. Lower retained error must consequently be reported with actual retention, subgroup composition and review workload. The crop–season calibration extension targets the observed coverage deficit but has not yet been shown to correct it. Point estimates and interval endpoints are expressed on the log1p scale before inverse transformation. The inverse-transformed mean log prediction is generally not the arithmetic conditional mean of productivity. Consequently, original-scale point predictions, interval endpoints and log-scale error measure different aspects of the predictive distribution; improved RMSLE alone does not establish unbiased original-scale mean prediction.

4.6 Phenology attribution and climate ablation

Table 12. Reported absolute SHAP magnitudes and climate-block ablations.

Evidence type	Feature / configuration	Value
SHAP 1	Corrected minimum relative humidity	0.0857
SHAP 2	Average temperature	0.0585
SHAP 3	Post-transplant rainfall	0.0567
SHAP 4	Pre-to-transplant radiation change	0.0545
SHAP 5	Average humidity	0.0454

SHAP 6	Temperature range	0.0327
Ablation	Context only	RMSLE 0.4438
Ablation	No post-transplant climate	RMSLE 0.4329
Ablation	No transplant climate	RMSLE 0.4320
Ablation	Full climate-enriched model	RMSLE 0.4305
Ablation	No pre-transplant climate	RMSLE 0.4300
Ablation	No seasonal climate	RMSLE 0.4296

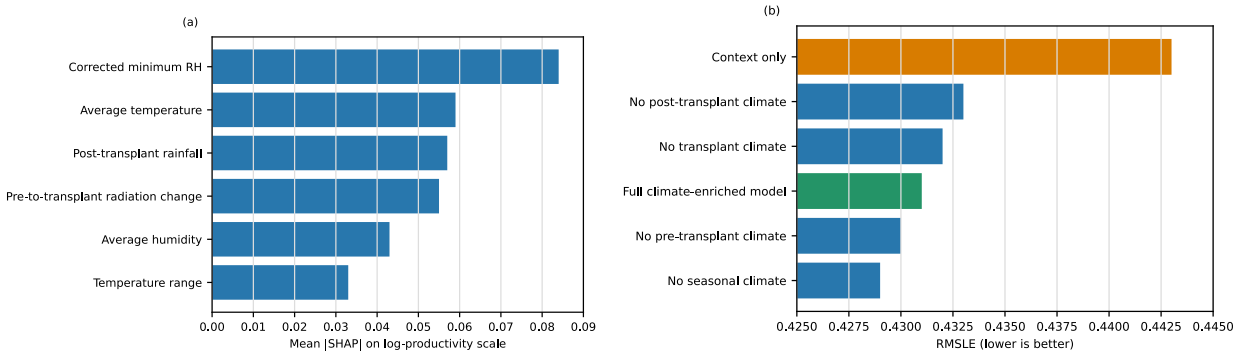


Figure 10. Feature attribution magnitude and RMSLE after climate-block removal.

Climate enrichment reduces RMSLE from 0.4438 to 0.4305, a difference of 0.0133 (3.0%) against context alone. Removing post-transplant climate produces the largest single-window deterioration, to 0.4329, but the relative change is only 0.56%. Removing transplant-month climate yields 0.4320. Individual-window effects are therefore modest compared with the joint climate contribution.

Removing pre-transplant or seasonal climate gives slightly lower errors of 0.4300 and 0.4296. The full feature set is consequently not the lowest-error configuration in the block comparison. Correlated exposures and estimation variability could produce these small reversals. Without paired uncertainty, the differences do not support a firm preference for excluding either block.

Corrected minimum relative humidity has the largest mean absolute SHAP magnitude, 0.0857, followed by average temperature (0.0585) and post-transplant rainfall (0.0567). The rainfall rank is consistent with the ablation ordering, but SHAP magnitude contains no sign information. It cannot establish whether wetter conditions increase productivity, identify a physiological optimum or quantify a causal response.

Attribution also depends on the explained model, output scale and background distribution. A TreeSHAP analysis of a boosting comparator cannot directly explain TabICLv2. Crop-specific signed dependence estimates for Aman, Aus and Boro, accompanied by uncertainty, would be needed for a crop-level interpretation. Unmeasured irrigation is particularly relevant to Boro. The present evidence supports phenology windows as useful predictors, while leaving crop-specific mechanisms unresolved.

4.7 Retrospective ranking and screening states

Table 13. Historical ranking agreement within district–season candidate sets.

Metric	CRPI only	Viability-weighted	Hurdle + uncertainty penalty
NDCG@3	0.673 ± 0.162	0.665 ± 0.159	0.615 ± 0.137
NDCG@5	0.699 ± 0.127	0.693 ± 0.132	0.654 ± 0.118

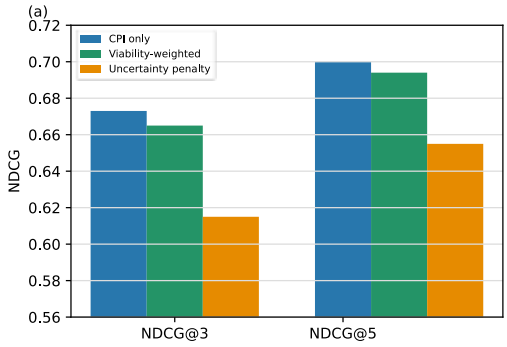


Figure 11. Ranking scores

CRPI-only ranking reaches NDCG@5 = 0.699. Viability weighting lowers it to 0.693, and uncertainty penalization reduces it further to 0.654. The total decline is 0.045 (6.4%); at K = 3 it is 0.058 (8.6%). Thus, the conservative score sacrifices agreement with historical productivity order. Any benefit from that sacrifice must be established through retained-set risk and the operational cost of abstention.

Table 14. Worked decision states using illustrative u_{40} = 1.0 and u_{60} = 2.0.

Crop	Prediction	Occurrence probability	Interval; uncertainty	Decision
Aus	2.00	0.90	[1.40, 2.60]; 0.53	Strong evidence / accept
Jute	3.00	0.65	[0.50, 5.50]; 1.54	Moderate evidence / review
Amra	2.00	0.80	[0.00, 7.00]; 3.11	ABSTAIN

The Bagerhat, Kharif-1 worked example uses average temperature 28°C, relative humidity 75% and transplant-month rainfall 180 mm as partial context. Table 16 demonstrates how the screening rule maps prediction summaries to decisions. With $\varkappa$ = 0.25, normalized uncertainties are 0.53, 1.54 and 3.11. Aus satisfies the strong-evidence rule; Jute satisfies the moderate-evidence rule; Amra exceeds u_{60} and abstains despite occurrence probability 0.80.

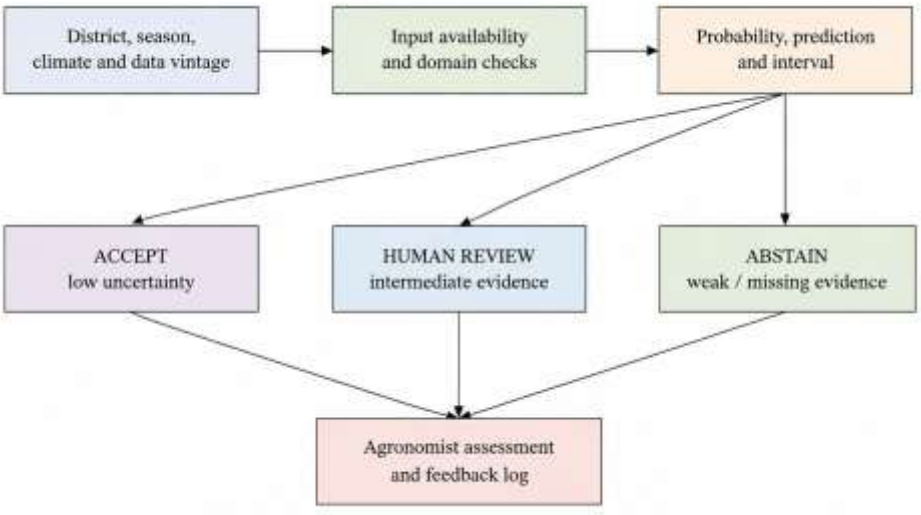


Figure 12. Decision routing after input and domain checks. Review remains distinct from automatic acceptance.

Historical candidates are the crops recorded in a district–season set, not all feasible alternatives. Rankings therefore cannot estimate the outcome of planting an unobserved crop. Missing soil, irrigation, cultivar and market information further restricts these outputs to screening for agronomic review.

4.8 Geographic validation scope

Geographic generalization was evaluated through district-grouped cross-validation across the available 64 districts. Each district was excluded from model fitting when assigned to an outer validation fold. Held-out districts therefore represent geographic separation within the study dataset, rather than additional districts from an external cohort. No separate external validation dataset was used. The results quantify transfer under this internal geographic partition; independent geographic and temporal validation remain future research objectives.

4.9 Scope of the empirical findings

District-grouped evaluation favours TabICLv2, while seasonal holdout, subgroup undercoverage and ranking degradation delimit that advantage. The following analyses examine paired model performance, component removal, uncertainty-based retention and fitted climate responses within the evaluated distribution.

4.10 Supplementary model comparisons and sensitivity analyses

All five models used identical test observations, preprocessing, feature representation and RMSLE calculation. District-level errors form 64 paired blocks. Let e_dj denote district d RMSLE for model j, calculated using Equation (6). Mean district RMSLE and the contrast against PHTF-Select are:

$$\bar{e}_j = \frac{1}{64}\sum_d e_{dj}\ ;\ \Delta_j = \frac{1}{64}\sum_d \left(e_{dj} - e_{d,PHTF}\right) \quad (7)$$

Complete district vectors were resampled 5,000 times to obtain bootstrap intervals while preserving model pairing. PHTF-Select has the lowest mean district RMSLE (0.3994), followed by SAINT (0.4035). Its relative reductions are 3.0% against FT-Transformer, 4.4% against TabTransformer, 1.0% against SAINT and 4.4% against NODE. These district-averaged estimates are distinct from the fold-averaged RMSLE in the main benchmark; 0.3994 is not substituted for 0.409.

Table 15. Mean district RMSLE for five models evaluated on common test data.

Model	Mean district RMSLE	95% bootstrap interval
PHTF-Select	0.3994	0.3827–0.4163
FT-Transformer	0.4116	0.3946–0.4284
TabTransformer	0.4180	0.4012–0.4347
SAINT	0.4035	0.3886–0.4191
NODE	0.4176	0.4011–0.4337

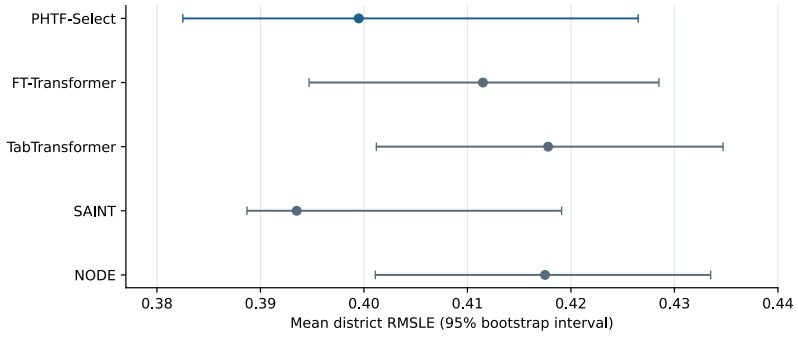


Figure 13. Mean district RMSLE and 95% bootstrap intervals on common test data. Lower values are better.

Table 16. Paired district RMSLE contrasts defined as comparator minus PHTF-Select.

Comparator	Mean RMSLE difference	95% bootstrap CI	Raw p	Holm p
FT-Transformer	0.0122	0.0025–0.0219	0.01759	0.03519
TabTransformer	0.0186	0.0100–0.0273	0.0002115	0.0006345
SAINT	0.0041	-0.0046–0.0129	0.4579	0.4579
NODE	0.0182	0.0108–0.0254	8.975e-06	3.59e-05

The Friedman test detects model differences, $\chi^2(4) = 27.4375$, $p = 1.621 \times 10^{-5}$. Holm-adjusted contrasts favour PHTF-Select over FT-Transformer ($p = 0.03519$), TabTransformer ($p = 0.0006345$) and NODE ($p = 3.59 \times 10^{-5}$). The SAINT contrast remains unresolved: 0.0041, CI [−0.0046, 0.0129], $p = 0.4579$. Thus, the lowest mean RMSLE does not establish superiority over every comparator. Inference follows paired contrasts in Table 18, not marginal-interval overlap.

4.10.1 Component ablation analysis

Table 17. Component ablations with complete retraining under a common evaluation protocol.

Configuration	Balanced accuracy	RMSLE	90% interval coverage	NDCG@5
Full framework	0.895	0.405	0.906	0.65
Without uncertainty	0.895	0.405	Undefined	0.69
Without hurdle	Undefined	0.433	0.891	0.631
Without phenology	0.872	0.431	0.895	0.622
Without crop-relative score	0.895	0.405	0.906	0.582
Without foundation model	0.878	0.439	0.902	0.619

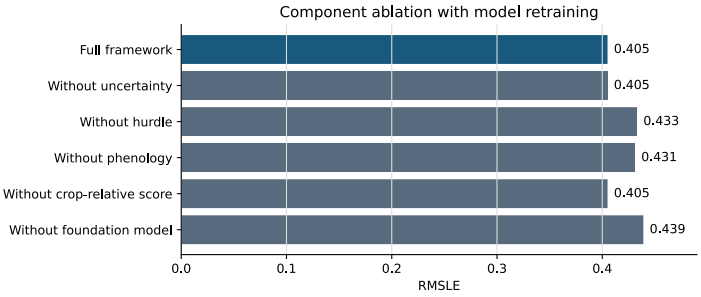


Figure 14. RMSLE after individual component removal and model retraining. The full-framework reference belongs to the separate ablation series.

Removing phenology increases RMSLE from 0.405 to 0.431 (+6.4%); removing the hurdle component increases it to 0.433 (+6.9%); replacing the foundation learner increases it to 0.439 (+8.4%). These controlled, retrained comparisons estimate each component's contribution conditional on the remaining framework. The 0.405 reference comes from the separate ablation execution, whereas 0.409 is the main benchmark result. Ablation effects are calculated against 0.405, without pooling the two executions.

Removing uncertainty leaves RMSLE unchanged at 0.405 but increases NDCG@5 from 0.650 to 0.690. Removing crop-relative scoring also leaves RMSLE unchanged while lowering NDCG@5 to 0.582, a 0.068 decline. Thus, these components primarily alter screening and ranking in this series. By contrast, removing phenology or the foundation learner affects both classification

and productivity. The ablations identify conditional contributions; without repeated-run intervals, small differences and component interactions remain unresolved.

4.10.2 Uncertainty and selective risk

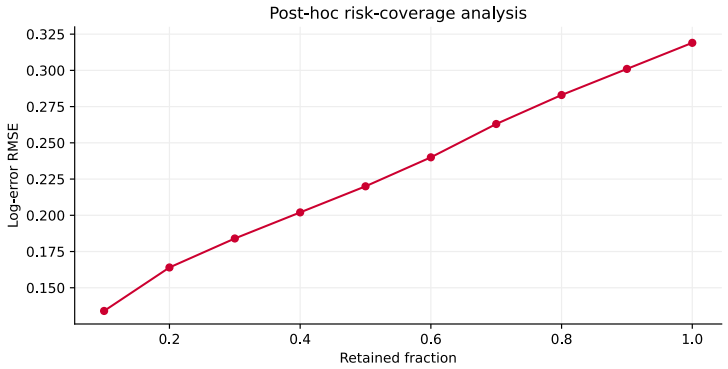


Figure 15. Post-hoc risk–coverage curve for 1,000 evaluated predictions ordered by normalized interval uncertainty. Risk is log-error RMSE.

Figure 15 uses 1,000 evaluated prediction cases and their interval-derived normalized uncertainty, as defined in Equation (3). Cases were ordered by uncertainty to examine error at successive retained fractions. This post-hoc analysis did not influence training, optimization or threshold selection. Log-error RMSE decreases from 0.3192 at full retention to 0.2398 at 60% retention (24.9%). This subset-specific result is separate from the main RMSLE reduction from 0.409 to 0.321 (21.5%; Table 13). The curve supports uncertainty-based error ordering within these cases, without establishing a universal uncertainty–error relationship.

4.10.3 Climate response analysis

$$f_{ris}(r, t, h) = 1 + 0.12r - 0.06r^2 - 0.10t^2 - 0.04h. \quad (8)$$

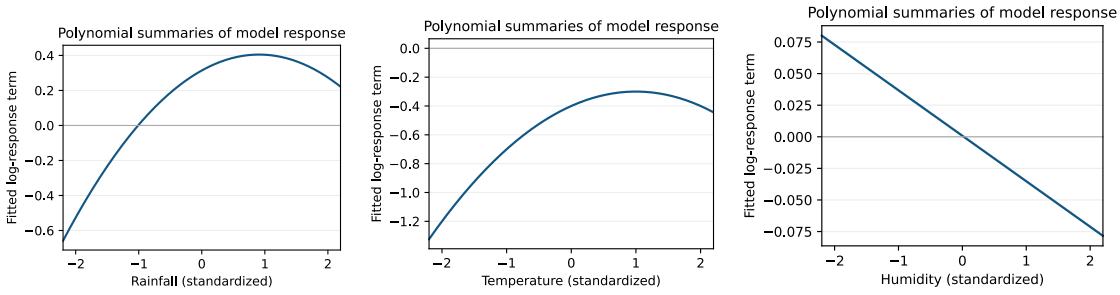


Figure 16. Polynomial summaries of trained-model climate responses. Visualization fits are separate from the predictor.

Trained-model responses were examined against rainfall, temperature and humidity with reference to the observed feature distribution. Equation (8) summarizes these responses for visualization. At (r, t, h) = (1, 1, 1), fitted contributions sum to −0.08; the rainfall term peaks at r = 1. These features characterize the visualization fit, not a physiological optimum or causal effect. The additive summary does not establish the predictive model's interactions.

4.11 Discussion

PHTF-Select improved both stages of the screening problem under complete district holdout. TabICLv2 increased balanced accuracy by 0.009 and AUROC by 0.005 over TabPFN v2, while its regression RMSLE was 0.008 lower. The margins are modest at the fold level, but the district-paired analysis is more informative: the 0.021 RMSLE reduction relative to XGBoost corresponds to a 4.9% relative improvement, the bootstrap interval excluded zero, and 73% of districts favoured TabICLv2. These results support a consistent advantage across locations rather than a gain driven by one favourable fold.

The viability task was strongly imbalanced, with 90.69% positive records. AUROC alone would therefore overstate operational quality. TabICLv2 combined the highest MCC (0.621) with the lowest Brier score (0.046) and ECE (0.019), indicating that its advantage extended beyond ranking to threshold-relevant probability quality. Even so, the label represents recorded co-occurrence of positive area and production. It should not be interpreted as a biological suitability endpoint because zero values may also reflect reporting practice, market choice, or management constraints.

The leakage control produced the largest performance change in the study. Restoring Area and Production reduced RMSLE from 0.430 to 0.104, an apparent 75.8% improvement created by algebraic access to the outcome. This result is not a competitive benchmark; it quantifies the bias that arises when components of Production/Area are treated as predictors. The smaller reduction from 0.430 to 0.414 under shuffled K-fold validation shows a second source of optimism. Records from related district-season environments become mixed across folds, whereas district holdout evaluates transfer to an unseen administrative unit.

Seasonal transfer exposed a substantially stronger distribution shift. Leave-one-season-out RMSLE ranged from 0.600 to 0.893, and most R^2 estimates were non-positive. Thus, the model interpolates across districts within represented seasons more reliably than it extrapolates to a season excluded from training. This distinction is central to deployment: a geographically new district within a known seasonal regime is not equivalent to a new climatic regime. The abstention rule appropriately rejects an unseen crop-season state, but rejection does not solve the underlying lack of evidence.

Phenology-window variables produced a measurable but limited gain. The full climate-enriched model improved RMSLE by about 3% relative to context-only features, and removing post-transplant climate caused the largest single-block deterioration. Post-transplant rainfall also ranked third by mean absolute SHAP value. In contrast, removing seasonal or pre-transplant blocks slightly improved RMSLE, consistent with redundancy among correlated climate summaries. The attribution values quantify predictive reliance, not response direction or causality; crop-specific signed analyses are required before agronomic interpretation.

The importance of post-transplant information also creates a decision-timing constraint. Those observations are unavailable when a farmer chooses what to plant. The current model is therefore best understood as retrospective or in-season screening. A pre-plant system would need weather forecasts or climatological substitutes and must be revalidated at a fixed decision date. Using realized post-plant weather in a pre-plant recommendation would introduce temporal information that is unavailable in practice.

Uncertainty evaluation revealed a second aggregate-to-subgroup gap. Overall interval coverage was close to the nominal 90% level, yet coverage fell to 79.8% in the highest-productivity quartile. At 60% retention, TabICLv2 reduced RMSLE from 0.409 to 0.321, a 21.5% relative reduction, but abstained on 40% of records. The operating point is useful only when the cost of review or non-decision is lower than the expected cost of acting on a high-error estimate. Coverage, interval width, retained-set error, and abstention rate must therefore be reported together.

Retrospective ranking reinforces this multi-objective interpretation. CRPI-only ranking achieved the highest NDCG@5 (0.699), while viability weighting reduced it to 0.693 and the uncertainty penalty to 0.654. The penalty intentionally moves uncertain candidates downward, so reduced agreement with historical productivity order is expected. Evaluation should compare ranking quality at matched retention and, in prospective work, incorporate decision costs rather than treating NDCG as the sole measure of utility.

The contribution of PHTF-Select lies in the integration of a leakage-resistant two-stage target, phenology-aligned predictors, district-held-out benchmarking, crop-relative ranking, predictive intervals, and an explicit abstention policy. This integration addresses several weaknesses identified in crop-yield modelling reviews [30,45,51]. The framework does not establish a universally superior tabular learner, and its results should not be transferred beyond the observed data domain without new calibration and validation.

4.11.1 Implications for Agricultural Decision Support

A practical implementation should display the viability probability, productivity estimate, prediction interval, uncertainty score, and acceptance state together. Agronomists should review cases with wide intervals, weak viability evidence, missing inputs, or crop-season combinations absent from training. Soil, irrigation, cultivar, fertilizer, pest pressure, labour, and market information

must be added before the output can support planting decisions. The present evidence supports prioritizing candidates for review, not replacing local agronomic judgement.

4.12 Limitations

The study has six material limitations. First, AgriClimateBD is cross-sectional and lacks row-level year identifiers, so district holdout does not test future-year forecasting. Second, a common physical productivity unit across all 72 crops was not independently established; CRPI reduces scale sensitivity but does not replace unit harmonization. Third, the maximum productivity value of 951.217 indicates extreme skew and requires source-level verification. Fourth, district grouping does not enforce separation by adjacency or agroecological zone. Fifth, crop occurrence is observational and cannot identify causal suitability. Sixth, climate summaries omit soil, irrigation, cultivar, management, pest, labour, and economic variables. These constraints limit interpretation to within-dataset screening under represented seasons.

Future validation should add multiple years, independent data sources, agroecological blocking, pre-plant feature availability, crop-specific attribution, and prospective evaluation with extension specialists. Calibration should be examined by crop, season, productivity range, and district group. Releasing the cleaned records, fold map, model configurations, out-of-fold predictions, interval outputs, and analysis scripts would allow independent reproduction.

5. Conclusion

PHTF-Select separates recorded cultivation viability from conditional productivity and evaluates both tasks under district-held-out validation. TabICLv2 achieved an AUROC of 0.956, balanced accuracy of 0.895, and regression RMSLE of 0.409. Its district-paired RMSLE was lower than XGBoost in 73% of districts. Uncertainty-based selection reduced RMSLE to 0.321 at 60% retention, although interval coverage fell below nominal levels for high-productivity records. Phenology features improved accuracy, but their post-transplant availability restricts the framework to retrospective or in-season use unless forecast inputs are validated. The evidence supports uncertainty-aware crop screening within represented conditions; autonomous crop recommendation requires temporal external validation and additional agronomic and economic variables.

ACKNOWLEDGEMENTS

The authors sincerely acknowledge Dr. A. O. M. Shamsuddoha, University of Liberal Arts Bangladesh (ULAB), for his valuable supervision, academic guidance, constructive suggestions, and continuous support throughout the development of this research.

The authors also acknowledge the Bangladesh Bureau of Statistics (BBS), AgriClimateBD dataset developers, and NASA POWER services for providing the agricultural and climate data resources used in this study. The authors appreciate the support and contributions of colleagues and the open-source scientific community whose tools and resources facilitated the analysis and implementation of the proposed framework.

AUTHOR CONTRIBUTIONS

Sadikul Hasan Mridha Atul: Conceptualization, Methodology, Software, Formal Analysis, Visualization, Writing – Original Draft, Project Administration.

Md. Isthiakul Hasan: Software, Data Curation, Investigation, Validation, Writing – Review & Editing.

Sayed Easin Arafat: Methodology, Formal Analysis, Validation, Writing – Review & Editing.

Md Shahriar Islam: Formal Analysis, Validation, Visualization, Writing – Review & Editing.

Mahmudul Hasan: Validation, Formal Analysis, Resources, Writing – Review & Editing.

Abu Obayed Alfatha: Data Curation, Software, Investigation, Writing – Review & Editing.

Farhana Sultana: Formal Analysis, Resources, Writing – Review & Editing.

A. O. M. Shamsuddoha: Supervision, Writing – Review & Editing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

The study involved secondary analysis of publicly available agricultural and climate datasets and did not include human participants, animal experiments, or personal data collection. Therefore, ethical approval and informed consent were not applicable for this research.

DATA AVAILABILITY STATEMENT

The data used in this study are publicly available. AgriClimateBD Version 2 can be accessed through the Mendeley Data repository: doi:10.17632/tb5x9gw2pt.2.

REFERENCES

[1] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy, and S.-H. Kim, "Random forests for global and regional crop yield predictions," PLOS ONE, vol. 11, no. 6, Art. e0156571, 2016, doi: 10.1371/journal.pone.0156571.

[2] J. 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.

[3] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in Proc. Thirty-First AAAI Conf. Artificial Intelligence, vol. 31, no. 1, pp. 4559–4566, 2017, doi: 10.1609/aaai.v31i1.11172.

[4] F. M. Talaat, "Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes," Neural Computing and Applications, vol. 35, pp. 17281–17292, 2023, doi: 10.1007/s00521-023-08619-5.

[5] D. Paudel, A. J. W. de Wit, H. Boogaard, D. Marcos, S. Osinga, and I. N. Athanasiadis, "Interpretability of deep learning models for crop yield forecasting," Computers and Electronics in Agriculture, vol. 206, Art. 107663, 2023, doi: 10.1016/j.compag.2023.107663.

[6] T. Islam, T. Mazumder, M. N. S. Roni, and M. S. Nur, "A comparative study of machine learning models for predicting Aman rice yields in Bangladesh," Heliyon, vol. 10, no. 23, Art. e40764, 2024, doi: 10.1016/j.heliyon.2024.e40764.

[7] M. K. Majumder, M. S. Rahman, R. K. Mondal, and M. S. Akter, "Climate-smart agriculture and food security in climate-vulnerable coastal areas of Bangladesh," Heliyon, vol. 10, no. 22, Art. e39885, 2024, doi: 10.1016/j.heliyon.2024.e39885.

[8] J. Biswas, S. Kumar, and M. S. Basunia, "AgriClimateBD: A satellite enriched precision agriculture dataset for Bangladesh," Mendeley Data, Version 2, 2026, doi: 10.17632/tb5x9gw2pt.2.

[9] R. Chowdhury, F. N. Nur, M. N. Islam, Md. N. Islam, P. Das, and A. S. Afridi, "SPAS-Dataset-BD: Dataset for smart precision agriculture system in Bangladesh," Data in Brief, vol. 61, Art. 111727, 2025, doi: 10.1016/j.dib.2025.111727.

[10] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," Nature, vol. 637, pp. 319–326, 2025, doi: 10.1038/s41586-024-08328-6.

[11] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICL: A tabular foundation model for in-context learning on large data," in Proc. 42nd Int. Conf. Machine Learning, PMLR 267, pp. 50817–50847, 2025.

[12] Y. Gorishniy, A. Kotelnikov, and A. Babenko, "TabM: Advancing tabular deep learning with parameter-efficient ensembling," in Proc. Int. Conf. Learning Representations, 2025.

[13] I. Rubachev, N. Kartashev, Y. Gorishniy, and A. Babenko, "TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks," in Proc. Int. Conf. Learning Representations, 2025.

[14] Y. Zeng, T. Dinh, W. Kang, and A. C. Mueller, "TabFlex: Scaling tabular learning to millions with linear attention," in Proc. 42nd Int. Conf. Machine Learning, PMLR 267, pp. 74051–74079, 2025.

[15] J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan, "TabICLv2: A better, faster, scalable, and open tabular foundation model," arXiv:2602.11139, 2026.

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

[17] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30, pp. 3146–3154, 2017.

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems 31, 2018.

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

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
[21] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in Advances in Neural Information Processing Systems 9, pp. 155–161, 1997.
[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, pp. 4765–4774, 2017.
[23] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
[24] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," Journal of the Royal Statistical Society: Series B, vol. 82, no. 4, pp. 1059–1086, 2020, doi: 10.1111/rssb.12377.
[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, PMLR 70, pp. 1321–1330, 2017.
[26] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning, pp. 625–632, 2005, doi: 10.1145/1102351.1102430.
[27] 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.
[28] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," in Advances in Neural Information Processing Systems 32, pp. 3543–3553, 2019.
[29] V. Vovk, A. Gammerman, and C. Saunders, "Machine-learning applications of algorithmic randomness," in Proc. 16th Int. Conf. Machine Learning, pp. 444–453, 1999.
[30] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
[31] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, "blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models," Methods in Ecology and Evolution, vol. 10, no. 2, pp. 225–232, 2019, doi: 10.1111/2041-210X.13107.
[32] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.
[33] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979.
[34] C. Nadeau and Y. Bengio, "Inference for the generalization error," Machine Learning, vol. 52, no. 3, pp. 239–281, 2003, doi: 10.1023/A:1024068626366.
[35] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," ACM Transactions on Information Systems, vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.
[36] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems 30, pp. 4878–4887, 2017.
[37] J. G. Cragg, "Some statistical models for limited dependent variables with application to the demand for durable goods," Econometrica, vol. 39, no. 5, pp. 829–844, 1971, doi: 10.2307/1909582.
[38] 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.
[39] Bangladesh Bureau of Statistics, Yearbook of Agricultural Statistics of Bangladesh 2022. Dhaka: Statistics and Informatics Division, Ministry of Planning, 2023.
[40] NASA Langley Research Center, "Prediction Of Worldwide Energy Resources (POWER): Solar and meteorological data services," accessed July 2026.
[41] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.
[42] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.
[43] 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.

[44] E. Elbasi, C. Zaki, A. E. Topcu, W. Abdelbaki, A. I. Zreikat, E. Cina, A. Shdefat, and L. Saker, "Crop prediction model using machine learning algorithms," Applied Sciences, vol. 13, no. 16, Art. 9288, 2023, doi: 10.3390/app13169288.

[45] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in Advances in Neural Information Processing Systems 35, 2022.

[46] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems 34, pp. 18932–18943, 2021.

[47] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in Proc. 35th AAAI Conf. Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687, 2021, doi: 10.1609/aaai.v35i8.16826.

[48] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv:2012.06678, 2020. https://arxiv.org/abs/2012.06678.

[49] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," arXiv:2106.01342, 2021. https://arxiv.org/abs/2106.01342.

[50] S. Popov, S. Morozov, and A. Babenko, "Neural oblivious decision ensembles for deep learning on tabular data," International Conference on Learning Representations, 2020. https://openreview.net/forum?id=r1eiu2VtwH.

[51] 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.

[52] R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano, "Split conformal prediction and non-exchangeable data," Journal of Machine Learning Research, vol. 25, no. 225, pp. 1–38, 2024. https://jmlr.org/papers/v25/23-1553.html.