Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 09 May 2026
Revised 01 June 2026
Accepted 12 June 2026

Natural Resources for Human Health 2026, Vol. 6 (11s): 975–996

KEYWORDS:
Rice yield forecasting; XGBoost; Long Short-Term Memory (LSTM); TreeSHAP; Chronological validation; Climate risk screening; Indian agriculture

Explainable Multimodal Artificial Intelligence Framework for Climate Resilient Crop Yield Prediction and Sustainable Farm Decision Support

[1]Kadar Shereef I, [2]Angeline Christobel Y, [3]Nandhini J, [4]Lijetha C Jaffrin

[1]Assistant Professor, Department of Computer Science, The New College, Affiliated to Madras University, Chennai, Tamil Nadu, India, kadarshereef@thenewcollege.edu.in
[2]Associate Professor, Department of Computer Science, Faculty of Liberal Arts and Business Studies, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu, India, angelinechristobel5@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, RMD Engineering College, Affiliated to Anna University, Thiruvallur, Tamil Nadu, India, Nandhini.raman20@gmail.com
[4]Assistant Professor (Senior Grade) Department of Information Technology, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India, drlijethacjaffrin@veltech.edu.in

ABSTRACT: Forecasting rice yield at the district level is central to agricultural planning and contingency management in India, yet the task is complicated by spatial heterogeneity, weather volatility, and strong temporal persistence in production. This study builds a provenance-audited, leakage-free predictive framework that pairs an XGBoost branch handling static and lagged agronomic predictors with a five-year Long Short-Term Memory (LSTM) network capturing sequential weather dynamics. Using a filtered multi-decadal panel of 300 Indian districts spanning 1981–2017, models were trained on data through 2012, tuned on 2013–2014, and independently evaluated on 703 chronologically held-out records from 2015–2017. A validation-selected late-fusion scheme (0.80 XGBoost, 0.20 LSTM) reached an RMSE of 477.29 kg/ha and an R2 of 0.743, meaningfully ahead of standalone XGBoost (485.87 kg/ha) and standalone LSTM (512.26 kg/ha). Paired statistical testing, however, found no significant edge over strong tabular baselines—Ridge Regression (477.44 kg/ha) and Random Forest (477.47 kg/ha). TreeSHAP attributions showed that predictions are driven mainly by historical yield persistence, secular year trends, and rainfall rather than intricate temporal patterns. Stratified diagnostics further revealed sharp performance loss under rainfall deficit (RMSE 547.71 kg/ha) and compound heat-dryness stress (560.93 kg/ha), together with pronounced disparities across macro-regions. Taken together, the results suggest that although hybrid architectures offer useful diagnostic interpretability, any operational deployment must account for persistence bias, regional variance, and calibrated uncertainty during climatic extremes.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

1.1 Background

Rice (Oryza sativa) underpins food security, rural livelihoods, and macroeconomic stability throughout India [1], [5]. Cultivation stretches across sharply different agro-ecological settings—canal-fed plains, rainfed lowlands, coastal floodplains, and marginal upland tracts [1], [12]—so that national and state-level averages tend to obscure large swings in outcomes between and within districts [5], [12]. Yield itself emerges from the interplay of crop genetics, soil physicochemical conditions, agronomic management, and weather occurring at sensitive phenological windows [1], [8], [15], and none of these drivers acts in isolation [12], [26]. Reliable irrigation can cushion a region against a weak monsoon, while purely rainfed systems remain highly exposed to moisture stress [12]; likewise, how efficiently nitrogen fertilizer is taken up depends on soil moisture, and heat anomalies do differing amounts of damage depending on whether they strike during vegetative growth, panicle initiation, or grain filling [5], [8]. District-level rice yield modelling is therefore a multi-factorial, nonlinear, and spatio-temporally entangled problem that resists simple extrapolation from national trends [12], [26].

1.2 Importance of Research Problems

As climate volatility accelerates, estimating yield at the district level has moved from a descriptive exercise to an operational necessity for agricultural planning [12], [28]. Changes in the Indian summer monsoon, rising average and extreme temperatures, erratic rainfall, and compound heat-moisture stress are all pushing up production variance [12], [26]. The Intergovernmental Panel on Climate Change (IPCC) argues that continental or global averages cannot by themselves guide adaptation policy—resilience planning needs district-specific risk intelligence [12].

India's official post-harvest assessment still relies on crop-cutting experiments [5]; while these remain the authoritative benchmark, finalized figures are typically released months after harvest [17]. Data-driven forecasting can complement official statistics by flagging likely production shortfalls early and identifying districts that may need agronomic support [17], [28], which in turn helps with market interventions, food-grain procurement, distribution planning, and crop-insurance settlement [24], [29]. But such tools cannot operate as uncalibrated black boxes [18], [21]—for real adoption, a model must expose its own weaknesses, quantify uncertainty, and be transparent about which predictors are driving its output [18], [24].

1.3 Limitations of Existing Solutions

Machine learning applications in agriculture have grown quickly [14], [17], [29], but several structural gaps persist when these methods are pushed toward operational, district-level panels:

- **Methodological Divide Between Tabular and Sequence Learners:** Tree ensembles such as Random Forest [2], [13] and regularized gradient boosting (XGBoost) [4] handle nonlinear interactions and irregular tabular thresholds well but treat successive years as independent unless lag features are engineered by hand [4]. Recurrent architectures such as LSTM [9], [15] capture sequential dependence directly, but they typically need long, continuous sequences and can underperform or overfit on small, low-frequency annual panels dominated by strong autoregressive signal [3], [14].
- **Inadequate Benchmark Scrutiny:** Deep architectures are frequently benchmarked only against trivial baselines [14], [29], leaving open whether they genuinely beat well-tuned regularized linear models such as Ridge Regression or standard tree ensembles [2], [6], [26].
- **Data Provenance Blind Spots:** Multi-decadal public weather archives often blend real sensor readings with simulated climatological backfills or long-term normals to plug missing values [20]. Models trained on such unvetted panels risk learning synthetic artifacts rather than genuine weather–yield relationships [17], [29].
- **Data Leakage via Random Splitting:** A common flaw in crop-yield studies is random train–test partitioning [3], [24], [29]. Because district-year records are spatially and temporally correlated, random splits scatter neighbouring years and districts across both sets, so the resulting metric measures interpolation rather than genuine forward forecasting [3], [24].

- **Opaque and Unverified Attributions:** Explainable-AI tools such as SHAP [18], [19] are increasingly used, but their attributions are often read as causal agronomic effects, obscuring how heavily models actually lean on spatial proxies and historical persistence rather than dynamic weather signals [18], [21].

1.4 Research Gap Identification

Despite considerable progress in remote sensing and deep learning [10], [15], [25], [30], no single district-scale framework yet combines all of the following:

1. Provenance-audited data filtering that removes climatological backfills [11], [20];
2. Leakage-free, multi-year chronological holdout validation [3], [24];
3. A principled hybrid architecture joining structured gradient boosting with recurrent temporal learning [4], [9], [31];
4. Rigorous paired significance testing against strong, classical tabular baselines [2], [6]; and
5. Diagnostic, non-causal explainability paired with stress-stratified error evaluation under compound climatic shocks [12], [18], [19].

Much of the existing work either evaluates on field-scale imagery or short time spans that do not scale nationally [10], [25], [30], or operates at a broad macro level without testing whether the extra algorithmic complexity actually beats regularized regression [5], [26], [29].

1.5 Objectives

The central aim of this study is to design, implement, and evaluate a validation-weighted hybrid XGBoost–LSTM architecture for district-level Kharif rice yield forecasting across India, and to establish whether fusing a structured branch with a recurrent branch produces a verifiable gain over either component model and over standard baselines [4], [9], [31].

The supporting objectives are to:

- Quantify the incremental value of distinct feature modalities—historical yield lags, weather variables, soil proxies, and agricultural inputs—through structured ablation experiments [23], [26], [29];
- Assess how reliably the hybrid model forecasts future years across India's Central, East, North, South, and West macro-regions [5], [12];
- Stress-test performance under adverse weather, specifically rainfall deficit, excess precipitation, extreme heat, and compound heat-dryness anomalies [8], [12], [26]; and
- Use TreeSHAP [18], [19] to inspect model behaviour, isolate the leading feature attributions, gauge the model's reliance on historical yield persistence, and mark the boundary between predictive risk screening and causal agronomic prescription [18], [21].

1.6 Major Contributions

This study advances district-level agricultural yield forecasting along four fronts:

- **Provenance-Audited, Leakage-Free Dataset Construction:** A rice-specific panel covering 300 districts across 20 Indian states (1981–2017) was assembled from ICRISAT and NASA POWER sources [11], [20], with strict provenance filtering that removed climatological temperature backfills and 30-year normal rainfall substitutions. Evaluation follows a chronological split (training through 2012, validation across 2013–2014, and an independent test set of 703 sequence-eligible records spanning 2015–2017), with all transformations fitted only on historical data [3], [24].
- **Principled Hybrid XGBoost–LSTM Late-Fusion Architecture:** A modular design routes static agronomic covariates, soil characteristics, and explicit yield lags through an XGBoost branch [4], while ordered multi-variable sequences pass through a five-year LSTM branch [9]. The fusion weight ($\alpha = 0.80$ XGBoost, 0.20 LSTM) was chosen solely on the validation split, keeping evaluation on future horizons unbiased [3], [31].

- **Methodologically Honest and Statistically Audited Benchmarking:** The hybrid reached an independent test RMSE of 477.29 kg/ha and an R^2 of 0.743, improving significantly on standalone XGBoost ($p = 0.00043$) and standalone LSTM ($p < 0.000001$) under paired Wilcoxon signed-rank tests [6]. The manuscript is explicit that fusion did not significantly outperform strong conventional baselines (Ridge Regression: 477.44 kg/ha, $p = 0.237$; Random Forest: 477.47 kg/ha, $p = 0.918$) [2], [6], indicating that historical autoregression dominates annual district panels.
- **Comprehensive Stress-Stratification and TreeSHAP Diagnostics:** TreeSHAP attributions [18], [19] show that three-year rolling yield lags and temporal trend indicators supply the main predictive signal, with annual climate contributing secondary refinement. Subgroup analysis exposed clear performance drops under rainfall deficit (RMSE 547.71 kg/ha) and compound heat-dryness (RMSE 560.93 kg/ha), plus large regional gaps (Northern RMSE 286.73 kg/ha versus Central RMSE 789.98 kg/ha) [12], pointing to concrete criteria for operational risk screening.

2. LITERATURE REVIEW

2.1 Review of the Models

Agricultural yield forecasting has moved from classical parametric techniques toward nonlinear ensembles, deep recurrent models, and multi-source hybrids [12], [29].

- **Parametric and Regularized Linear Models:** Classical yield modelling relied on ordinary least squares, which assumes linear relationships and independent observations [5], [26]. Once multi-collinearity appears among precipitation, temperature, and fertilizer inputs, plain linear models suffer from inflated variance [5], [26]. Penalized regression—Ridge Regression with L_2 regularization—addresses this by shrinking coefficients and stabilizing estimation across correlated predictors [5], [26]. Empirical work on Indian rice yield finds that regularized linear baselines stay competitive whenever the panel carries a strong persistent autoregressive signal [5], [26].
- **Tree-Based Ensemble Methods:** Random Forest [2], [13] and XGBoost [4] have become standard benchmarks for tabular agricultural modelling [17], [29]. Random Forest builds parallel trees via bootstrap aggregation and random subspace sampling, damping variance and tolerating collinear covariates [2], [13]. XGBoost instead minimizes a gradient-based loss with explicit depth and weight penalties (L_1 and L_2 regularization) to control overfitting [4]. Both capture arbitrary thresholds and nonlinear interactions—for instance, the joint effect of high temperature and soil moisture deficit—without a pre-specified functional form [4], [13]; however, in their native form they treat each district-year as an independent observation, with no internal state to represent ordered temporal dependence [4].
- **Recurrent Sequence Models:** Recurrent networks pass hidden states across time steps to represent ordered dynamics [9]. The LSTM architecture adds input, forget, and output gates plus a dedicated cell state to counter vanishing and exploding gradients [9]. In crop forecasting, LSTMs can capture multi-year carryover effects such as soil moisture depletion and cumulative input history [8], [15]. Even so, deep sequence learners face real constraints on annual district tables, where sequences are short, sample sizes are bounded, and inputs are low-frequency administrative summaries rather than continuous phenological curves [8], [14].
- **Hybrid and Multi-Source Frameworks:** Hybrid learning combines structured tabular representations with recurrent or spatial ones [27], [31]. When the component models make largely independent errors, late-fusion ensembles can in principle beat either branch alone [31]. But in agricultural panels where historical persistence dominates total variance, the structured and temporal branches often pick up redundant signal [27], so hybrids need validation-weighted fusion and rigorous benchmarking against strong tabular baselines before any complexity gain can be taken as genuine [3], [6], [31].

2.2 Comparative Analysis of Prior Work

Prior work spans very different operational scales, data inputs, and objectives [14], [17], [29]. At the field and sub-district scale, several recent studies lean on high-resolution remote sensing to track localized crop vigor [10], [21], [25], [30]. UAV-based multispectral imagery combined with tree ensembles has achieved high precision for rice grain yield by directly observing canopy reflectance and panicle development [25], [30], and Sentinel-2 time series have similarly been used to extract vegetation and water indices for rice yield prediction [10], [21]. Remote sensing captures within-season physiological responses that coarse

climate variables miss, but field- and drone-based studies do not scale easily to nationwide historical district panels, given persistent cloud cover during the Kharif monsoon, sensor turnover, and the difficulty of aggregating plot-level rasters to administrative boundaries [10], [25]. At the regional and national scale, studies more often run machine learning over weather and soil databases [1], [8], [15]. Work that integrates crop genetics, downscaled meteorological projections, and deep architectures underscores the value of heterogeneous inputs [8], [15], and Indian comparisons of linear, neural, and penalized regression across agro-climatic zones confirm that meteorological variables carry real predictive power [5], [26]. Yet several comparative analyses note that the apparent edge of complex models often comes down to which baselines were chosen [6], [29]—many studies compare advanced neural networks only against a persistence mean or an unregularized linear model, leaving out tuned gradient boosting or regularized Ridge regression [5], [26], [29].

2.3 Limitations and Research Opportunities

A critical read of the prior literature surfaces three recurring methodological gaps:

- **Evaluation Leakage via Random Partitioning:** Random train-test splits remain common in crop-yield literature [3], [24], [29]. Because agricultural records show strong spatial proximity and multi-year autocorrelation, random partitioning places records from the same district and adjacent years into both training and test sets [3], [24]. This measures interpolation rather than genuine future-year forecasting and artificially deflates error metrics [3], [24]; chronological holdouts are needed to mimic real operational conditions [3], [24].
- **Provenance Gaps in Public Environmental Panels:** Large-scale agricultural studies often ingest open-access meteorological products without careful auditing [11], [20]. Public archives frequently fill historical gaps with long-term 30-year normals or simulated climatological backfills [20]. When such artificial records are fed into machine learning pipelines, models learn static averages rather than true year-to-year variation, quietly injecting bias into the resulting predictions [17], [29].
- **Uncritical Explainability:** Explainable-AI tools—particularly SHAP [18] and TreeSHAP [19]—are being adopted more widely, but attributions are frequently read as causal agronomic mechanisms [21]. In practice, tree-ensemble feature importance often latches onto geographic identifiers and autoregressive yield lags, which act as proxies for unmeasured infrastructure, irrigation access, or reporting practices rather than direct biophysical drivers [18], [19], [21].

These gaps together motivate an auditable, provenance-filtered district-scale panel, evaluated on a multi-year chronological holdout, and subjected to paired statistical testing alongside diagnostic, non-causal explainability [3], [6], [18], [24].

2.4 Research Justification

Administrative crop planning, buffer-stock procurement, and relief programs in India all need dependable, auditable district-scale risk screening [12], [28]. Field-scale experimental plots capture fine-grained agronomic processes [10], [25], but macro-level planning needs multi-decadal coverage across hundreds of administrative units [11], [12].

Building an explainable hybrid that couples XGBoost's structured feature mapping [4] with an LSTM's sequential memory [9] is justified precisely because it lets us test, rather than assume, whether temporal deep learning adds marginal value over tabular baselines in annual panels [8], [31]. A strict chronological evaluation horizon (training through 2012, validation across 2013–2014, testing across 2015–2017) further gives an honest read on future-year forecasting accuracy, so that reported performance reflects what an operational deployment could actually expect [3], [24].

2.5 Related Works and Literature Survey Table

Table 1 summarizes a comparative survey of 11 related studies spanning crop-yield prediction methods and evaluation strategies.

Natural Resources for Human Health
eISSN: 2583-1194

Table 1: Comparative literature survey of recent studies

Authors and Year	Data and Domain	Model or Approach	Finding and Limitation
Antony and Karuppasamy, 2023 [1]	Indian paddy soil data	Machine-learning models	Shows soil variables can support paddy productivity prediction; local data limit transfer.
Ha et al., 2023 [8]	Korean rice and meteorology	SVM and LSTM	Tests observed and downscaled weather; geography and climate products affect generalization.
Htun et al., 2023 [10]	Sentinel-2 rice fields	Vegetation indices and ML	Demonstrates spectral value; image availability and field aggregation constrain scale.
Khaki et al., 2021 [15]	Genotype and weather	Deep neural network	Integrates heterogeneous data; genotype information is unavailable in many district panels.
Newman and Furbank, 2021 [21]	Satellite crop traits	Explainable machine learning	Connects traits and predictions; explanations remain associative.
Sarkar et al., 2024 [25]	UAV rice imagery	Ensemble Learning	Improves field-scale prediction; UAV coverage limits historical national use.
Setiya et al., 2023 [26]	Indian rice weather data	Regression and neural models	Provides weather-based comparison; limited spatial and temporal scope.
Shahhosseini et al., 2021 [27]	US Corn Belt	Crop model plus machine learning	Hybrid process-data modelling improves prediction; crop and region differ from Indian rice.
Talaat, 2023 [28]	IoT precision agriculture	Crop Yield Prediction Algorithm	Links sensing and decision support; requires operational IoT infrastructure.
Zhao et al., 2023 [31]	Chinese rice multi-source data	Hybrid LSSVM	Shows benefit of hybrid multi-source learning; station-scale design limits national transfer.
Van Klompenburg et al., 2020 [29]	Crop-yield studies	Systematic review	Identifies common features and algorithms; highlights inconsistent validation practice.

Natural Resources for Human Health
eISSN: 2583-1194

3. PROBLEM FORMULATION

3.1 Mathematical Formulation

Let $i \in \{1, 2, \ldots, N\}$ index administrative districts and $t \in \{1, 2, \ldots, T\}$ index agricultural crop years [11]. The target response $y_{it} \in \mathbb{R}^+$ denotes the Kharif rice yield observed in district i during harvest year t, in kilograms per hectare (kg/ha) [11]. The task is to learn a mapping $\mathcal{F}$ that projects predictors available up to time t to an accurate estimate $\hat{y}_{it}$, subject to a strict temporal non-anticipation constraint—no observation from a year $\tau > t$ may be used [3], [24]. The overall framework splits information into a structured tabular branch and a sequential temporal branch [4], [9]:

Structured Tabular Branch (f_X): The structured predictor vector x_{it} combines contemporaneous weather observations, static soil proxies, agricultural input quantities, administrative location identifiers, and engineered autoregressive yield lags [11], [20], [23]:

$$x_{it} = [\, w_{it}^T,\ s_i^T,\ a_{it}^T,\ g_i^T,\ y_{i,t-1},\ y_{i,t-2},\ y_{i,t-3},\ \bar{y}_{i,t}^{(3)} \,]^T$$

where w_{it} comprises annual mean temperature, cumulative precipitation, and relative humidity [20]; s_i denotes categorical soil orders and representative soil pH [23]; a_{it} represents chemical fertilizer consumption (nitrogen, phosphorus, potassium, and total macronutrients) alongside pesticide applications [11]; g_i captures one-hot encoded state and district geographic vectors [11]; and $\bar{y}_{i,t}^{(3)} = (1/3)\Sigma_{k=1}^{3} y_{i,t-k}$ is the three-year moving yield average. The structured model $f_X(x_{it}; \theta_X)$ is instantiated as an Extreme Gradient Boosting (XGBoost) ensemble of M regression trees [4]:

$$\hat{y}_{X,it} = \Sigma_{m=1}^{M} f_m(x_{it}),\ f_m \in \mathcal{H}$$

optimized by minimizing a regularized objective combining prediction loss and tree-complexity penalties [4]:

$$\mathcal{L}_X(\theta_X) = \Sigma_{i,t}\, l(y_{it}, \hat{y}_{X,it}) + \Sigma_{m=1}^{M} \Omega(f_m),\ \ \Omega(f_m) = \gamma T_m + \tfrac{1}{2}\lambda \|w_m\|^2$$

where l is the squared-error loss, T_m is the number of terminal leaf nodes, and w_m is the leaf weight vector [4].

Temporal Recurrent Branch (f_L): The temporal sequence S_{it} aggregates L = 5 consecutive annual observations of time-varying environmental and agronomic covariates preceding and including year t [9], [15]:

$$S_{it} = [\, z_{i,t-4},\ z_{i,t-3},\ z_{i,t-2},\ z_{i,t-1},\ z_{it} \,]^T$$

where each step vector z_{it} contains temperature, precipitation, humidity, soil pH, nitrogen, phosphorus, potassium, total fertilizer, pesticide consumption, and the previous-year yield $y_{i,t-1}$ [11], [20]. The sequence is processed by an LSTM through recursive input, forget, and output gates that update hidden states h_τ and cell states c_τ [9]:

$$f_\tau = \sigma(W_f z_{i\tau} + U_f h_{\tau-1} + b_f)$$

$$i_\tau = \sigma(W_i z_{i\tau} + U_i h_{\tau-1} + b_i)$$

$$\tilde{c}_\tau = \tanh(W_c z_{i\tau} + U_c h_{\tau-1} + b_c)$$

$$c_\tau = f_\tau \odot c_{\tau-1} + i_\tau \odot \tilde{c}_\tau$$

$$o_\tau = \sigma(W_o z_{i\tau} + U_o h_{\tau-1} + b_o)$$

$$h_\tau = o_\tau \odot \tanh(c_\tau)$$

A dense feed-forward layer then maps the final hidden state h_t to a scalar prediction: $\hat{y}_{L,it} = W_d h_t + b_d$ [9].

Validation-Weighted Late Fusion: The two branch predictions are combined via a convex combination [27], [31]:

$$\hat{y}_{F,it} = a\hat{y}_{X,it} + (1 - a)\hat{y}_{L,it},\ \ a \in [0,1]$$

The fusion coefficient α is tuned strictly on the validation partition by minimizing RMSE across a candidate grid spanning zero to one [3]:

$$a^* = \operatorname{argmin}_{a \in [0,1]} \sqrt{\,[\, (1/|D_{val}|)\, \Sigma_{(i,t) \in D\{subval\}} \{ y_{it} - [a\hat{y}_{X,it} + (1-a)\hat{y}_{L,it}] \}^2 \,]}$$

Natural Resources for Human Health
eISSN: 2583-1194

The validation minimum falls at $\alpha^* = 0.80$, assigning 80% weight to the structured XGBoost branch and 20% to the sequential LSTM branch [4], [9], [31].

3.2 Research Assumptions

The framework's validity rests on four explicit assumptions [3], [11], [20], [24]:

1. **Spatial Unit Harmonization:** District-year records are assumed comparable over time once harmonized to stable historical district boundaries, so that boundary splits and administrative reorganizations do not distort the multi-decadal series [11].
2. **Availability of Lagged Production Statistics:** Yields for $t-1$, $t-2$, and $t-3$ are assumed finalized and available at prediction time, acting as non-causal statistical proxies for persistent local irrigation, infrastructure, and baseline productivity [5], [26].
3. **Strict Temporal Isolation of Transformations:** All preprocessing—imputation medians, scaling parameters, and categorical encoding vocabularies—is fitted only on training years ($t \leq 2012$), eliminating look-ahead leakage [3], [24].
4. **Data Provenance Authenticity:** Retained climate fields are assumed to be authentic annual meteorological records, verified by removing climatological backfills and static 30-year normal rainfall substitutions from the public source databases [20].

3.3 Dataset Characteristics

The source data integrates the ICRISAT District Level Database, NASA POWER agro-climatology services, and FAOSTAT agricultural input statistics [11], [20]. After provenance filtering and lag generation, the executable panel contains 9,213 rice records covering 300 districts in 20 Indian states over 1981–2017 [11]. Enforcing an uninterrupted five-year sequence requirement produces a common modelling sample, partitioned chronologically as follows [3], [24]:

- **Training Partition (1984–2012):** 6,668 sequence-eligible records (out of 7,965 total eligible tabular observations), used for model training and parameter estimation.
- **Validation Partition (2013–2014):** 531 sequence-eligible records (out of 535 tabular records), reserved strictly for hyperparameter tuning and late-fusion weight selection.
- **Independent Test Partition (2015–2017):** 703 sequence-eligible records (out of 713 tabular records), forming an independent, out-of-sample future horizon for evaluation.

The input modalities differ in spatial resolution [11], [20], [23]:

- **District-Year Specific:** Kharif rice yield, cumulative precipitation, mean surface temperature, relative humidity, nitrogen, phosphorus, potassium, and aggregate fertilizer consumption all vary by district and year [11], [20].
- **Static Proxies and Macro Aggregates:** Soil pH is a soil-order cross-walk proxy [23], while pesticide consumption enters as a national annual time series because disaggregated district-level history is unavailable [11].

3.4 Hypothesis Development

Five formal hypotheses are tested on the common 703 independent test records to verify the framework's operational utility, baseline competitiveness, and robustness [3], [6]:

- **Hypothesis 1 (H_1 – Component Improvement):** Validation-weighted late fusion yields significantly lower RMSE and absolute error on unseen future crop years than either the standalone XGBoost [4] or standalone LSTM [9] branch.
- **Hypothesis 2 (H_2 – Baseline Superiority):** The XGBoost-LSTM hybrid significantly outperforms strong conventional tabular baselines, namely regularized Ridge Regression [5], [26] and Random Forest [2], [13].

Natural Resources for Human Health
eISSN: 2583-1194

- **Hypothesis 3 (H_3 – Incremental Modality Value):** Progressively adding dynamic climate variables, soil proxies, and agricultural inputs to historical yield lags produces measurable error reduction relative to a history-only baseline [26].
- **Hypothesis 4 (H_4 – Climate Stress Degradation):** Prediction error rises under adverse meteorological anomalies specifically severe rainfall deficit and compound heat-dryness stress relative to performance under normal conditions [12].
- **Hypothesis 5 (H_5 – Spatial Generalization Disparity):** Predictive accuracy varies materially across India's Central, East, North, South, and West macro-regions, reflecting differences in irrigation cover, soil type, and climate exposure [12].

4. RESEARCH WORK

4.1 Data Pipeline and Provenance Filtering

The data assembly pipeline was built around three priorities: multi-decadal temporal depth, administrative continuity, and empirical auditability. The primary multi-source panel cross-references district-level production statistics from the ICRISAT District Level Database [11], meteorological services from NASA POWER [20], and agro-chemical input data linked through FAOSTAT [11]. The raw database initially spanned 1966 through 2017 across the major rice-producing administrative units.

A provenance audit of the public meteorological records surfaced two quality issues that needed correcting:

- Temperature and relative humidity values prior to 1981 were climatological backfills rather than direct daily satellite or ground measurements.
- Rainfall series in several districts substituted long-term 30-year station normals during reporting gaps.

Because ingesting simulated averages creates artificial stability and misleading learning signal, a provenance-filtering protocol removed all records before 1981 along with any observation derived from climatological-normal substitution. The usable panel therefore begins in 1981, and because the structured branch needs a three-year historical window to compute lags, the effective target horizon starts in 1984.

4.2 Leakage-Aware Preprocessing and Feature Modalities

Feature preprocessing followed a strictly leakage-free design throughout. Median imputation for occasional missing values, one-hot encoding dictionaries for categorical variables, and standard scaling were all calibrated solely on the training partition (1984–2012); the resulting parameters were then applied, without re-estimation, to freeze-transform the validation (2013–2014) and independent test (2015–2017) partitions. Engineered features fall into five groups:

- **Historical Yield Dynamics:** single-year (t−1), two-year (t−2), and three-year (t−3) yield lags, plus a rolling three-year moving average ($\bar{y}_{i,t}^{(3)}$) computed within each district panel.
- **Dynamic Agro-Climatic Features:** annual cumulative precipitation, mean surface temperature, and relative humidity from NASA POWER [20].
- **Soil Characteristics:** harmonized categorical soil orders and a representative soil pH proxy reflecting regional edaphic profiles [23].
- **Agricultural Inputs:** district-year consumption of nitrogen (N), phosphorus (P), potassium (K), and total chemical fertilizer, plus national-scale pesticide application rates [11].
- **Technological and Spatial Proxies:** calendar-year indices capturing secular technological progress, together with one-hot encoded state and district identifiers absorbing persistent unmeasured spatial variation.

4.3 Algorithmic Baseline Hierarchy and Architectural Design

To check whether deep sequential architectures offer measurable gains over simpler statistical approaches, a multi-tier baseline hierarchy was built:

Natural Resources for Human Health
eISSN: 2583-1194

1. **District Historical Mean:** a non-parametric persistence baseline predicting future yield as the historical district training mean.
2. **Ridge Regression:** an L_2-penalized linear model over scaled static and lagged features, testing linear predictability under multicollinearity [5], [26].
3. **Random Forest:** a non-parametric ensemble of randomized decision trees over structured tabular covariates [2], [13].
4. **Standalone Structured Branch (XGBoost):** an Extreme Gradient Boosting model over tabular covariates and autoregressive lags [4]. The configuration selected via validation RMSE used 700 trees, a maximum depth of 4, a learning rate (η) of 0.03, minimum child weight of 2, row subsampling of 0.85, and column subsampling of 0.85.
5. **Standalone Temporal Branch (LSTM):** a recurrent network taking five consecutive annual steps (L = 5) of 10 dynamic environmental and agronomic covariates [9]. The optimized configuration used 32 hidden LSTM units, a dropout rate of 0.15, a 32-unit dense feed-forward layer, Adam optimization [16], and early stopping with weight restoration at epoch 78.
6. **Validation-Weighted Late Fusion:** a convex ensemble combining the structured prediction $\hat{y}_X$ and the temporal prediction $\hat{y}_L$: $\hat{y}_{F,it} = \alpha\hat{y}_{X,it} + (1-\alpha)\hat{y}_{L,it}$. The scalar α was swept over [0, 1] in steps of 0.05 on the validation split; minimum validation RMSE occurred at $\alpha = 0.80$, allocating 80% weight to XGBoost and 20% to LSTM.

4.4 End-to-End Methodological Workflow

Figure 1 illustrates the overall end-to-end framework, from provenance extraction through to statistical risk auditing.

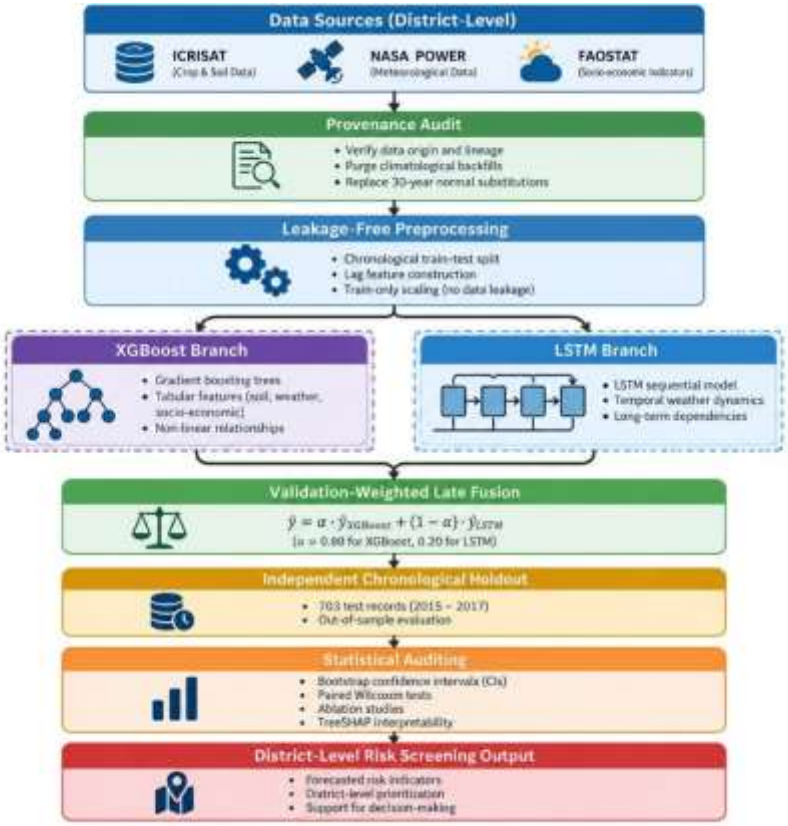

Figure 1: Workflow of the proposed provenance-aware hybrid framework.

The formal algorithmic sequence governing dataset filtering, parameter estimation, model training, and diagnostic evaluation is set out in Algorithm 1.

Natural Resources for Human Health
eISSN: 2583-1194

ALGORITHM 1: HYBRID TRAINING AND EVALUATION PROCEDURE

1. Ingest the provenance-enriched district-level agricultural panel and isolate rice records.
2. Purge records prior to 1981, remove climatological backfills, and drop normal-rainfall substitutions.
3. Sort observations chronologically by district and compute three autoregressive lags ($y_{i,t-1}$, $y_{i,t-2}$, $y_{i,t-3}$) and rolling three-year means ($\bar{y}_{i,t}^{(3)}$).
4. Partition records chronologically into training (1984–2012), validation (2013–2014), and independent testing (2015–2017) subsets.
5. Fit imputation, one-hot encoding, and scaling operators strictly on the training partition; apply the frozen transformers to validation and test subsets.
6. Train the persistence mean, Ridge, Random Forest, and candidate XGBoost configurations; select optimal XGBoost hyperparameters using validation RMSE.
7. Construct five-year rolling sequence matrices (S_{it}) and train the LSTM network with early stopping.
8. Evaluate late-fusion weights $\alpha \in [0, 1]$ on the validation partition to establish the optimal convex parameter α^*.
9. Freeze all model architectures and parameters; generate predictions for the 703 common sequence-eligible test observations.
10. Compute global loss metrics (MAE, RMSE, MAPE, R^2), derive 95% district-block bootstrap confidence intervals, run paired Wilcoxon signed-rank tests, perform feature ablations, assess regional/weather subgroup disparities, and generate TreeSHAP explanations.

4.5 Experimental Dataset Configurations

Enforcing the five-year continuous history requirement across all comparative models produced a strictly identical evaluation sample. Table 2 details the partition boundaries and resulting sequence counts.

Table 2: Dataset partitions

Partition	Chronological Horizon	Eligible Tabular Records	Five-Year Sequences (N)
Training	1984–2012	7,965	6,668
Validation	2013–2014	535	531
Independent Test	2015–2017	713	703

The complete taxonomy of baseline models, candidate learners, and the late-fusion framework evaluated across the 703 test records is summarized in Table 3.

Table 3: Executed modelling configuration

Model	Input Feature Set	Algorithmic Role
Historical Mean	District production records	Historical persistence benchmark
Ridge Regression	Scaled static features and yield lags	Regularized linear baseline [5], [26]

Natural Resources for Human Health
eISSN: 2583-1194

Model	Input Feature Set	Algorithmic Role
Random Forest	Structured district-year features and lags	Non-parametric ensemble baseline [2], [13]
XGBoost	Tabular multi-source features and lags	Standalone structured branch [4]
LSTM	5-year ordered sequences of 10 covariates	Standalone temporal recurrent branch [9]
Proposed Fusion	Validation-weighted outputs ($\alpha = 0.80$)	Integrated late-fusion hybrid

4.6 Statistical Testing and Diagnostic Verification Protocol

Evaluation relied on four standardized loss metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2)—computed over the n = 703 common independent holdout sample:

$$MAE = (1/n)\ \Sigma_{k=1}^{n}\ |y_k - \hat{y}_k|$$

$$RMSE = \sqrt{[\ (1/n)\ \Sigma_{k=1}^{n}\ (y_k - \hat{y}_k)^2\ ]}$$

$$MAPE = (100\%/n)\ \Sigma_{k=1}^{n}\ |(y_k - \hat{y}_k)\ /\ y_k|$$

$$R^2 = 1 - [\ \Sigma_{k=1}^{n}\ (y_k - \hat{y}_k)^2\ /\ \Sigma_{k=1}^{n}\ (y_k - \bar{y})^2\ ]$$

To verify statistical robustness without assuming normally distributed errors, two further diagnostic layers were run:

- **District-Block Bootstrap Resampling:** 1,000 bootstrap iterations resampled entire district clusters with replacement, preserving intra-district spatial and serial correlation, to generate an empirical 95% confidence interval for hybrid RMSE.
- **Paired Wilcoxon Signed-Rank Tests:** paired non-parametric tests checked whether absolute prediction-error differences between the hybrid model and each comparator baseline were stochastically significant [6].
- **TreeSHAP Attribution:** exact local feature attributions were computed across 500 representative test samples from the structured branch using TreeSHAP [19], decomposing predictions into additive feature importance values to separate historical persistence from dynamic weather sensitivity.

5. EXPERIMENTAL SETUP

5.1 Computing Environment and Reproducibility Controls

All computational workflows were implemented in Python using standard scientific computing, machine learning, and deep learning libraries:

- Data manipulation, feature engineering, and file I/O were handled with pandas and NumPy.
- Preprocessing scalers, encoders, and baseline regression models were built with scikit-learn [22].
- The gradient boosting architecture used the XGBoost library [4].
- Deep recurrent networks were built, trained, and run with TensorFlow Keras.
- Non-parametric statistical tests were implemented with SciPy.
- Feature attributions and additive explanations were computed with the SHAP package [18], [19].

Natural Resources for Human Health
eISSN: 2583-1194

The entire experiment was designed to run on CPU, since the five-year sequence dataset contains fewer than 7,000 training records and does not require GPU acceleration. To keep the study auditable and fully reproducible, deterministic random seeds were fixed across all library initializations—data splitting, model weight initialization, sequence generation, bootstrap iterations, and SHAP sample selection. Intermediate files, train–validation–test split indices, model definitions, and prediction arrays were saved as static artifacts rather than transcribed by hand.

5.2 Partitioning Design and Sequence Construction

Evaluation used a strict chronological split fixed before model fitting, to mirror real-world future-year forecasting:

- **Training Period (1984–2012):** 7,965 total eligible tabular observations, yielding 6,668 complete five-year sequence records.
- **Validation Period (2013–2014):** 535 total eligible tabular observations, yielding 531 complete five-year sequence records.
- **Independent Test Period (2015–2017):** 713 total eligible tabular observations, yielding 703 complete five-year sequence records.

The drop from raw eligible records to sequence-eligible records reflects sporadic historical discontinuities in individual district records, which occasionally prevented extraction of an uninterrupted five-year lookback window. Every comparative baseline, tree ensemble, and hybrid variant was evaluated on the exact same 703 independent test records from 2015–2017, so that this common-sample constraint rules out performance differences caused by sample-selection bias. Preprocessing parameters (median imputers, one-hot vocabularies, and standard scaling operators) were fitted exclusively on the 1984–2012 training subset and applied to later partitions without re-estimation.

5.3 Hyperparameter Configurations and Search Protocol

Hyperparameter search was restricted to bounded candidate sets on the 2013–2014 validation split, to avoid overfitting a small temporal validation window:

- **XGBoost Hyperparameters:** candidates varied tree count (300 to 1,000), maximum depth (3 to 6), and learning rate (0.01 to 0.1). The configuration selected by minimum validation RMSE used 700 estimators, a maximum depth of 4, a learning rate (η) of 0.03, minimum child weight of 2, row subsampling of 0.85, and column subsampling of 0.85.
- **LSTM Hyperparameters:** the sequence network processed 5 consecutive annual time steps with 10 features per step. The candidate grid covered hidden units (16 to 64) and dropout rates (0.1 to 0.3); the selected configuration used 32 hidden LSTM units, a dropout rate of 0.15, a 32-unit dense feed-forward layer, and a single scalar output node, optimized with Adam [16] on mini-batches. Overfitting was controlled with early stopping monitoring validation loss, restoring the optimal weights found at epoch 78.
- **Ridge Regression:** evaluated across a regularized L_2 penalty sweep on standardized features, with the regularization strength chosen on validation performance.
- **Random Forest:** implemented with standard ensemble hyperparameters and fixed random states.
- **Late-Fusion Tuning:** the fusion parameter α was swept over [0.0, 1.0] in steps of 0.05; the minimum validation RMSE occurred at $\alpha = 0.80$, fixing the hybrid weighting to 0.80 for the structured XGBoost prediction and 0.20 for the temporal LSTM prediction. The independent test partition remained sealed until every model parameter, stopping criterion, and fusion weight was locked in.

5.4 Evaluation Metrics, Uncertainty, and Subgroup Protocols

Model performance was tracked across the four metrics defined above—MAE, RMSE, MAPE, and R^2. RMSE served as the primary optimization and benchmarking criterion because it penalizes large administrative errors more heavily, while MAE and MAPE were tracked to confirm absolute and percentage error stability across the positive yield distribution. Diagnostic testing followed four standardized protocols:

- **District-Block Bootstrap Resampling:** to capture spatial and temporal autocorrelation within districts, 1,000 bootstrap iterations resampled entire district clusters with replacement, deriving an empirical 95% confidence interval for hybrid test RMSE.
- **Paired Non-Parametric Significance Testing:** paired Wilcoxon signed-rank tests were run on absolute prediction-error vectors between the late-fusion hybrid and each comparator model, testing the null hypothesis of symmetric error differences around zero.
- **Feature Modality Ablation:** the structured XGBoost pipeline was retrained on isolated and combined feature subsets: historical lags only, dynamic climate only, soil and agricultural inputs only, historical plus climate, and all static modalities combined.
- **Subgroup and Stress Diagnostics:** performance was audited across five geographic macro-regions (Central, East, North, South, West) and five meteorological anomaly regimes. Weather regimes (Normal, Rainfall Deficit, Excess Rainfall, High Temperature, and Combined Heat-Dryness) were defined using standardized z-scores derived strictly from each district's historical training distribution, preventing test-period information leakage.
- **Model Explainability:** TreeSHAP attributions were computed on a representative sample of 500 test records to quantify the mean absolute contribution of each tabular covariate on the final prediction scale.

6. RESULTS AND DISCUSSION

6.1 Independent Out-of-Sample Test Performance

The provenance-filtered panel yielded 703 sequence-eligible holdout observations across 300 districts for the independent evaluation window spanning 2015–2017. The proposed validation-weighted late-fusion architecture achieved an MAE of 336.51 kg/ha, an RMSE of 477.29 kg/ha, a MAPE of 21.35%, and an R^2 of 0.743—the lowest overall RMSE and highest explained variance among all evaluated models. Even so, the benchmark models were highly competitive. Tuned Random Forest produced the lowest MAE (329.45 kg/ha) and lowest MAPE (21.14%), with an RMSE of 477.47 kg/ha that is essentially indistinguishable from the hybrid. Regularized Ridge Regression reached an RMSE of 477.44 kg/ha—just 0.15 kg/ha above the fusion model—with an R^2 of 0.743. The historical-mean persistence benchmark performed worst on every metric (MAE 638.10 kg/ha, RMSE 805.78 kg/ha, MAPE 28.72%, $R^2 = 0.267$), confirming that static district averages cannot capture inter-annual yield volatility. Standalone XGBoost reached an RMSE of 485.87 kg/ha ($R^2 = 0.733$), and standalone LSTM registered an RMSE of 512.26 kg/ha ($R^2 = 0.704$). Late fusion reduced RMSE by 1.76% relative to standalone XGBoost and by 6.83% relative to standalone LSTM.

Table 4: Independent test performance

Model	MAE (kg/ha)	RMSE (kg/ha)	MAPE (%)	R^2
Proposed Fusion	336.51	477.29	21.35	0.743
Ridge Regression	337.94	477.44	21.61	0.743
Random Forest	329.45	477.47	21.14	0.743
XGBoost	343.65	485.87	22.08	0.733
LSTM	375.97	512.26	21.25	0.704
Historical Mean	638.10	805.78	28.72	0.267

Natural Resources for Human Health
eISSN: 2583-1194

Figure 2 illustrates the relative RMSE distributions across all candidate learners.

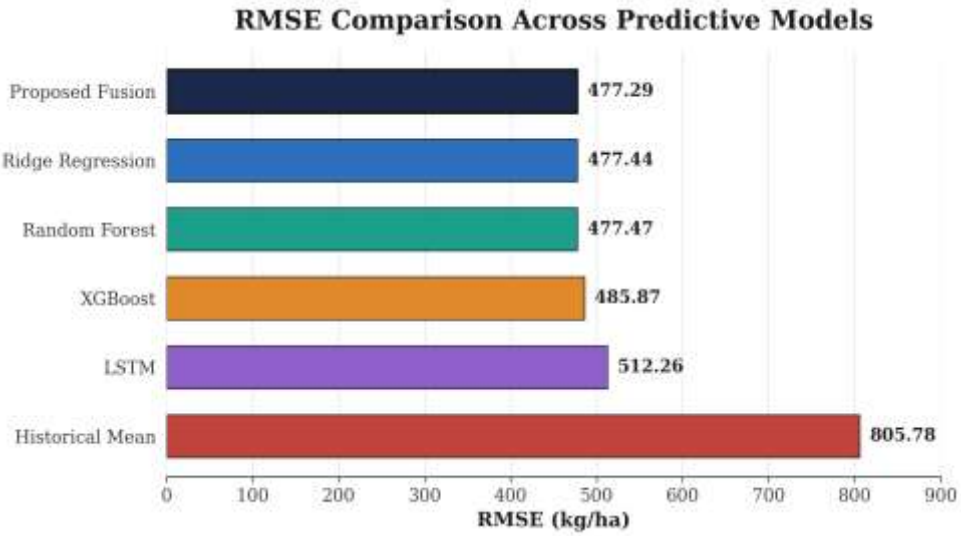

Figure 2: RMSE comparison across predictive models.

Plotting predicted against observed yields reveals broad alignment along the 1:1 identity line, alongside a clear tendency toward regression to the mean. Very high district yields (above 4,200 kg/ha) were consistently underpredicted, while several low-to-moderate observations were overpredicted—a pattern typical of squared-error optimization applied over heterogeneous spatial units.

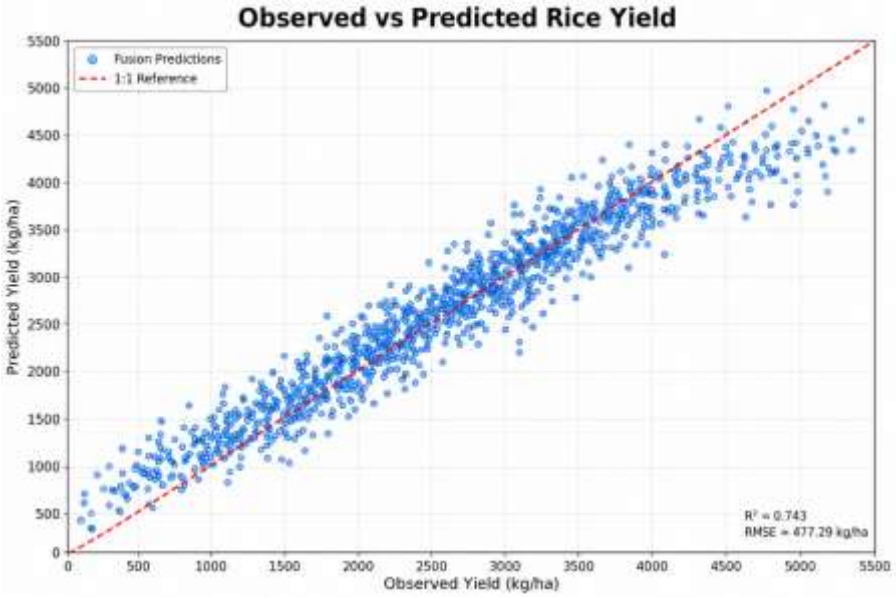

Figure 3: Observed and predicted rice yield for the proposed fusion model.

6.2 Statistical Hypothesis Testing (H_1 and H_2)

The empirical results support a nuanced reading of the two primary hypotheses:

- **Evaluation of Hypothesis 1 (H_1 – Component Improvement):** H_1 is supported. Paired Wilcoxon signed-rank tests confirmed that validation-weighted late fusion produced significantly lower absolute errors than standalone XGBoost ($p = 0.00043$) and standalone LSTM ($p < 0.000001$). District-block bootstrap resampling (1,000 replicates)

produced an empirical 95% confidence interval for hybrid RMSE spanning 428.28 to 528.78 kg/ha, quantifying the underlying spatial sampling uncertainty.

- **Evaluation of Hypothesis 2 (H_2 – Baseline Superiority):** H_2 is not supported. Paired absolute-error comparisons between the late-fusion hybrid and Ridge Regression showed no statistically detectable difference ($p = 0.237$), and the comparison against Random Forest likewise showed no significant separation ($p = 0.918$). In other words, in multi-decadal annual panels dominated by autoregressive persistence and structured indicators, adding a deep-sequence branch delivers only marginal gains over well-regularized classical baselines.

6.3 Feature Modality Ablation Analysis (H_3)

Ablation experiments on the structured branch isolated the predictive value of each data modality and support Hypothesis H_3. Historical yield features alone supplied the main predictive foundation, reaching an RMSE of 504.27 kg/ha ($R^2 = 0.713$). Using only dynamic climate features degraded performance substantially (RMSE 653.65 kg/ha, $R^2 = 0.518$), while soil and agricultural inputs alone gave an RMSE of 584.32 kg/ha ($R^2 = 0.615$). Combining historical lags with dynamic climate lowered RMSE to 493.87 kg/ha, and integrating all static modalities reached the final standalone structured performance of 485.87 kg/ha ($R^2 = 0.733$).

Table 5: XGBoost ablation results

Configuration	MAE (kg/ha)	RMSE (kg/ha)	MAPE (%)	R^2
Historical only	358.82	504.27	23.22	0.713
Climate only	488.72	653.65	30.07	0.518
Soil and inputs	439.95	584.32	27.00	0.615
History and climate	351.49	493.87	22.49	0.725
All static modalities	343.65	485.87	22.08	0.733

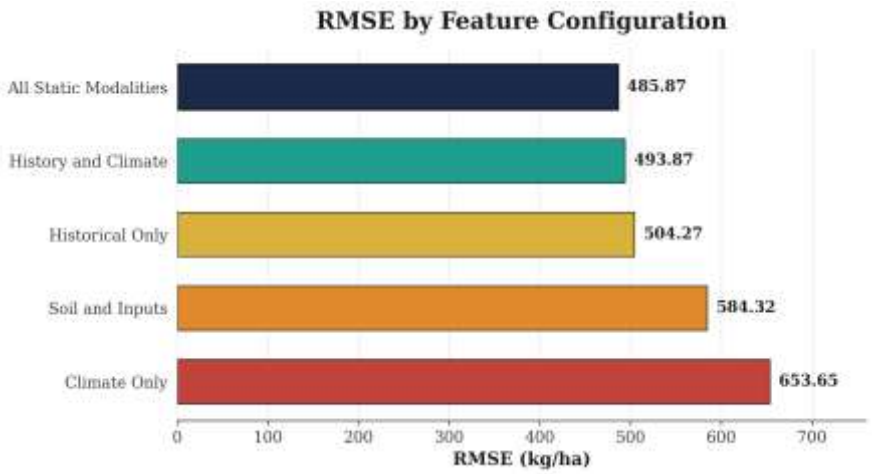

Figure 4: RMSE by feature configuration.

These results demonstrate that historical yield persistence functions as a powerful aggregated proxy for unmeasured local production capacity.

Natural Resources for Human Health
eISSN: 2583-1194

6.4 Regional Generalization Analysis (H_5)

Breaking model performance down by geographic macro-region fully supports Hypothesis H_5, exposing spatial disparities that the national aggregate RMSE conceals.

Table 6: Regional performance of the fusion model

Region	N	MAE (kg/ha)	RMSE (kg/ha)	MAPE (%)	R^2
North	187	211.49	286.73	9.89	0.910
West	86	302.55	380.89	29.00	0.766
South	171	315.79	408.18	12.68	0.738
East	159	358.69	510.28	22.73	0.440
Central	100	599.64	789.98	48.89	0.321

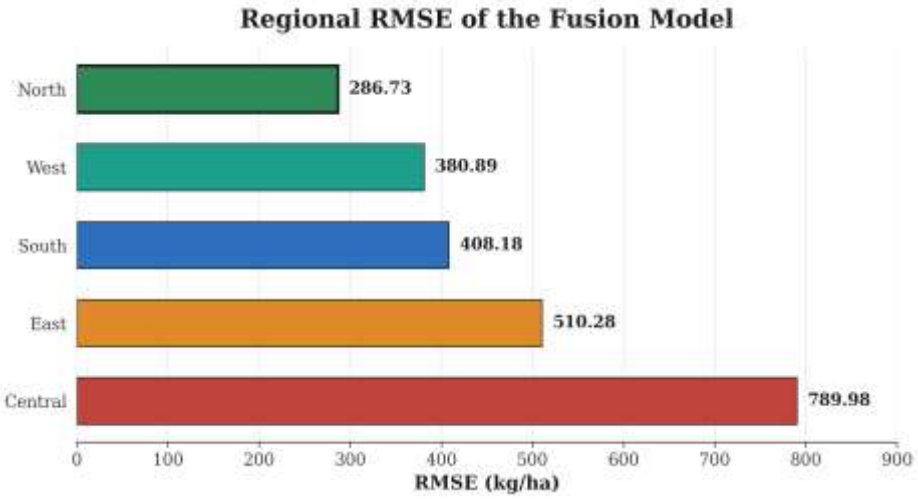

Figure 5: Regional RMSE of the fusion model.

Accuracy was highest in Northern districts (RMSE 286.73 kg/ha, MAPE 9.89%, $R^2 = 0.910$), where extensive canal and tubewell irrigation buffers crops against weather fluctuations. Reliability was markedly worse in Central districts (RMSE 789.98 kg/ha, MAPE 48.89%, $R^2 = 0.321$) and Eastern districts (RMSE 510.28 kg/ha, $R^2 = 0.440$), both dominated by rainfed cultivation, more complex edaphic regimes, and higher yield volatility. This indicates that any national-scale deployment of the model should be paired with localized uncertainty intervals.

6.5 Stress Diagnostics Under Climate Anomalies (H_4)

Stratifying test records by standardized historical meteorological anomaly regimes partially supports Hypothesis H_4. Under baseline "Normal" conditions (n = 326), the hybrid achieved an RMSE of 454.11 kg/ha ($R^2 = 0.795$). Performance degraded noticeably under adverse hydrological shocks:

- **Rainfall Deficit (n = 42):** RMSE rose to 547.71 kg/ha (MAPE 30.81%, $R^2 = 0.597$).

- **Combined Heat and Dryness (n = 88):** produced the highest error of any regime, at RMSE 560.93 kg/ha (MAPE 24.78%, R^2 = 0.624).
- **Excess Rainfall (n = 98):** showed elevated error (RMSE 509.83 kg/ha, R^2 = 0.708).
- **High Temperature Alone (n = 149):** RMSE of 426.97 kg/ha (R^2 = 0.682), actually below the normal baseline.

Table 7: Performance under normal and anomalous weather conditions

Condition	N	MAE (kg/ha)	RMSE (kg/ha)	MAPE (%)	R^2
High temperature	149	291.71	426.97	15.21	0.682
Normal	326	330.05	454.11	23.52	0.795
Excess rainfall	98	371.21	509.83	16.38	0.708
Rainfall deficit	42	401.86	547.71	30.81	0.597
Combined heat and dryness	88	366.42	560.93	24.78	0.624

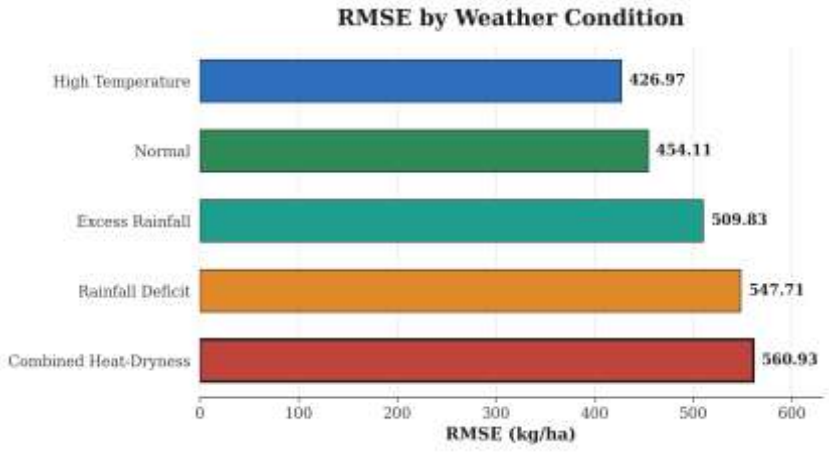

Figure 6: RMSE by weather condition**.**

The lower error observed under isolated high temperatures likely reflects the fact that, in the historical dataset, elevated temperatures frequently coincide with high-yielding, heavily irrigated northern basins where moisture stress is actively mitigated.

6.6 Model Interpretability via TreeSHAP

TreeSHAP feature attributions quantified the mean absolute contribution of each predictor on the model-output scale.

Table 8: Leading XGBoost features from TreeSHAP

Feature	Description	Mean Absolute SHAP Value
Yield_roll3	3-year rolling mean yield	509.50
Yield_lag1	Previous-year yield (t−1)	167.40

Natural Resources for Human Health
eISSN: 2583-1194

Feature	Description	Mean Absolute SHAP Value
Year	Technological trend counter	138.11
Yield_lag2	Two-year yield lag (t−2)	48.37
Rainfall	Annual cumulative precipitation	43.77
Yield_lag3	Three-year yield lag (t−3)	36.20
Nitrogen	Nitrogen fertilizer consumption	22.01
Fertilizer	Total macronutrient consumption	21.19
pH	Representative soil pH proxy	17.00
Humidity	Mean surface relative humidity	16.29
Phosphorus	Phosphorus fertilizer consumption	15.44
Temperature	Mean annual surface temperature	14.40

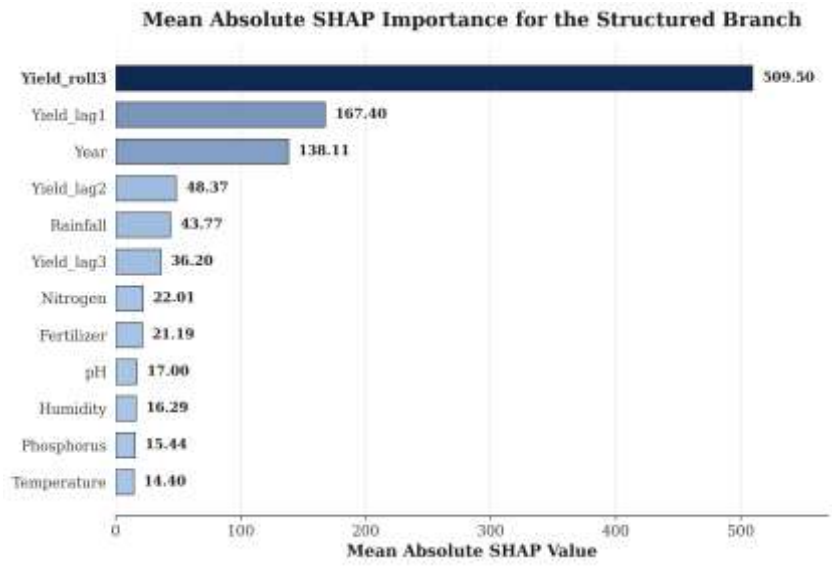

Figure 7: Mean absolute SHAP importance for the structured branch.

The rolling three-year average yield (Yield_roll3: 509.50) and the primary lag (Yield_lag1: 167.40) dominate the attributions, followed by the secular technological trend index (Year: 138.11). Among the dynamic environmental predictors, cumulative precipitation ranks highest (43.77), followed by nitrogen fertilizer use (22.01). These explanations confirm, diagnostically, that the model behaves fundamentally as a persistent yield-dynamics estimator, adjusted at the margin by rainfall anomalies and nutrient applications. These values represent predictive associations within an observational panel and should not be read as direct causal response coefficients for farm-level agronomic prescriptions.

7. CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

This study designed, implemented, and evaluated an explainable, provenance-audited hybrid framework for district-level Kharif rice yield forecasting across India. The architecture couples a structured Extreme Gradient Boosting (XGBoost) model with a

five-year sequential LSTM recurrent network through a validation-selected late-fusion scheme ($\alpha = 0.80$ XGBoost, 0.20 LSTM). The pipeline enforces leakage-free chronological partitioning (training: 1984–2012; validation: 2013–2014; independent testing: 2015–2017) across 300 districts in 20 states, and rigorously excludes climatological backfills and 30-year normal rainfall substitutions from the public source databases.

Across the 703 sequence-eligible independent holdout test records, the evaluation points to three core conclusions:

- **Component Optimization vs. Baseline Parity:** The late-fusion hybrid reached an RMSE of 477.29 kg/ha and an R^2 of 0.743, significantly outperforming standalone XGBoost (485.87 kg/ha, $p = 0.00043$) and standalone LSTM (512.26 kg/ha, $p < 0.000001$) under paired Wilcoxon signed-rank tests. Paired error comparisons, however, showed no statistically significant advantage over well-tuned classical baselines, including Ridge Regression (RMSE 477.44 kg/ha, $p = 0.237$) and Random Forest (RMSE 477.47 kg/ha, $p = 0.918$, lowest MAE of 329.45 kg/ha).
- **Dominance of Autoregressive Persistence:** Feature ablation and TreeSHAP attributions confirm that historical yield persistence (three-year rolling mean yield: 509.50; lag-1 yield: 167.40) and the secular technological trend index (138.11) supply most of the predictive power, with annual climate anomalies and chemical fertilizer inputs contributing incremental, marginal refinement rather than acting as primary drivers.
- **Vulnerability Under Climate Shocks and Regional Disparities:** Subgroup stress testing exposed real operational limits. Reliability fell substantially under hydrological shocks—rainfall deficit (RMSE 547.71 kg/ha) and compound heat-and-dryness stress (RMSE 560.93 kg/ha)—and national aggregate metrics masked steep regional disparities, from high precision in the heavily irrigated Northern plains (RMSE 286.73 kg/ha, $R^2 = 0.910$) to poor reliability in rainfed Central districts (RMSE 789.98 kg/ha, $R^2 = 0.321$).

In short, while the hybrid model offers an auditable, transparent framework for district-level risk screening, it should not be treated as an automated prescriptive engine or as an unconditional replacement for established tabular models and localized field assessment.

7.2 Limitations

Several empirical and structural constraints frame these findings:

- **Temporal Horizon of the Panel:** because of public data provenance exclusions and administrative release delays, the executable historical panel ends in 2017 and does not capture post-2017 climate extremes, cultivar shifts, or policy interventions.
- **Temporal Coarseness of Meteorological Covariates:** annual aggregate climate summaries smooth over short-duration, stage-critical agro-meteorological extremes—such as heat waves during anthesis or dry spells during panicle initiation—that actually drive biological yield failure.
- **Absence of High-Resolution Satellite Remote Sensing:** the long-term panel lacks continuous, historically harmonized satellite vegetation indices (e.g., MODIS NDVI/EVI or Sentinel-2 products) that would capture real-time canopy development and vegetative health.
- **Spatial Granularity and Proxy Inputs:** edaphic features rely on macro soil-order pH cross-walk proxies, pesticide application is represented as a national aggregate series, and fertilizer consumption reflects district totals rather than rice-specific field applications, which rules out farm-level causal inference.

7.3 Future Scope

Future work should push the operational viability and precision of agricultural yield forecasting along four technical dimensions:

- **Multi-Scale Remote Sensing and Phenological Alignment:** integrating cloud-masked MODIS and Sentinel-2 vegetation indices with within-season NASA POWER weather sequences aligned to localized crop-calendar stages (vegetative, flowering, grain-filling) to capture dynamic crop stress.

- **Architectural Advancements:** replacing the compact LSTM with Temporal Convolutional Networks (TCN) or Patch Time-Series Transformers (PatchTST) under the same leak-free chronological validation protocol, to test whether modern sequence models make better use of fine-grained weather time series.
- **Uncertainty Quantification and Spatial Calibration:** implementing conformal prediction or quantile gradient boosting to deliver mathematically calibrated prediction intervals for administrative planners, paired with leave-region-out retraining to evaluate spatial transfer to unobserved districts.
- **Decision-Centric and Economic Loss Evaluation:** moving beyond statistical loss metrics (RMSE/MAE) toward asymmetric, cost-sensitive utility functions and decision curves that explicitly weigh the economic cost of underpredicting severe production shortfalls against the cost of false-alarm interventions in procurement and crop-insurance operations.

REFERENCES

[1] Antony, A., & Karuppasamy, R. (2023). Mining of soil data for predicting the paddy productivity by machine learning techniques. Paddy and Water Environment, 21, 231-242. https://doi.org/10.1007/s10333-023-00924-y

[2] Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

[3] Cerqueira, V., Torgo, L., & Mozetic, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109, 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

[5] Das, B., Nair, B., Reddy, V. K., & Venkatesh, P. (2018). Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. International Journal of Biometeorology, 62, 1809-1822. https://doi.org/10.1007/s00484-018-1583-6

[6] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1-30.

[7] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. Advances in Neural Information Processing Systems, 9, 155-161.

[8] Ha, S., Kim, Y. T., Im, E. S., et al. (2023). Impacts of meteorological variables and machine learning algorithms on rice yield prediction in Korea. International Journal of Biometeorology, 67, 1825-1838. https://doi.org/10.1007/s00484-023-02544-x

[9] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Htun, A. M., Shamsuzzoha, M., & Ahamed, T. (2023). Rice yield prediction model using normalized vegetation and water indices from Sentinel-2A satellite imagery datasets. Asia-Pacific Journal of Regional Science, 7, 491-519. https://doi.org/10.1007/s41685-023-00299-2

[11] International Crops Research Institute for the Semi-Arid Tropics. (n.d.). District Level Database for Indian Agriculture. https://data.icrisat.org/dld/

[12] Intergovernmental Panel on Climate Change. (2022). Climate Change 2022: Impacts, Adaptation and Vulnerability. Cambridge University Press. https://www.ipcc.ch/report/ar6/wg2/

[13] Jeong, J. H., Resop, J. P., Mueller, N. D., et al. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

[14] Kamilaris, A., & Prenafeta-Boldu, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70-90. https://doi.org/10.1016/j.compag.2018.02.016

[15] Khaki, S., Pham, H., & Wang, L. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. PLOS ONE, 16(6), e0252402. https://doi.org/10.1371/journal.pone.0252402

[16] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations. https://arxiv.org/abs/1412.6980

[17] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

[18] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[19] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888. https://arxiv.org/abs/1802.03888

[20] NASA Langley Research Center. (n.d.). NASA POWER Data Services. https://power.larc.nasa.gov/docs/services/api/

[21] Newman, S. J., & Furbank, R. T. (2021). Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. Nature Plants, 7, 1354-1363. https://doi.org/10.1038/s41477-021-01001-0

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[23] Poggio, L., de Sousa, L. M., Batjes, N. H., et al. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. SOIL, 7, 217-240. https://doi.org/10.5194/soil-7-217-2021

[24] Roberts, D. R., Bahn, V., Ciuti, S., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929. https://doi.org/10.1111/ecog.02881

[25] Sarkar, T. K., Roy, D. K., Kang, Y. S., et al. (2024). Ensemble of machine learning algorithms for rice grain yield prediction using UAV-based remote sensing. Journal of Biosystems Engineering, 49, 1-19. https://doi.org/10.1007/s42853-023-00209-6

[26] Setiya, P., Satpathi, A., & Nain, A. S. (2023). Predicting rice yield based on weather variables using multiple linear, neural networks, and penalized regression models. Theoretical and Applied Climatology, 154, 365-375. https://doi.org/10.1007/s00704-023-04563-5

[27] Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports, 11, 1606. https://doi.org/10.1038/s41598-020-80820-1

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

[29] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

[30] Yang, Q., Shi, L., Han, J., Zha, Y., & Zhu, P. (2019). Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Research, 235, 142-153. https://doi.org/10.1016/j.fcr.2019.02.022

[31] Zhao, L., Qing, S., Wang, F., et al. (2023). Prediction of rice yield based on multi-source data and hybrid LSSVM algorithms in China. International Journal of Plant Production, 17, 693-713. https://doi.org/10.1007/s42106-023-00266-z