Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable and Leakage-Aware District-Level Wheat Yield Prediction Using Ensemble Learning, Recurrent Neural Networks, and a Lightweight Transformer: A Case Study of Madhya Pradesh, India

Viswavardhan Reddy Karna
viswavardhank@rvce.edu.in

R.V. College of Engineering

Neethu S
R.V. College of Engineering

Supreeth S
B.M.S. College of Engineering

Sarala D V
B.M.S. College of Engineering

Sunitha T
B.M.S. College of Engineering

Vishnu Vardhana Reddy Karna
SRM University, Andhra Pradesh

Article

Keywords: Wheat-yield prediction, leakage-aware validation, explainable machine learning, spatial–temporal generalisation, Sentinel-2, Madhya Pradesh

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10688441/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explainable and Leakage-Aware District-Level Wheat Yield Prediction Using Ensemble Learning, Recurrent Neural Networks, and a Lightweight Transformer: A Case Study of Madhya Pradesh, India

Viswavardhan Reddy Karna[1], Neethu S[2], Supreeth S[3], Sarala D V[4], Sunitha T[5], Vishnu Vardhana Reddy Karna[6]

viswavardhank@rvce.edu.in[1], neethus@rvce.edu.in[2], supreeth.cse@bmsce.ac.in[3], saraladv.cse@bmsce.ac.in[4], sunithat.mca@bmsce.ac.in[5], vishnuvardhanareddy.k@srmap.edu.in[6]

[1]Department of Artificial Intelligence and Machine Learning, RV College of Engineering, Bengaluru 560059, Karnataka, India

[2]Department of Computer Science and Engineering, RV College of Engineering, Bengaluru 560059, Karnataka, India

[3, 4]Department of Computer Science and Engineering, B.M.S. College of Engineering, Bengaluru 560004, Karnataka, India

[5]Department of Computer Applications, B.M.S. College of Engineering, Bengaluru 560004, Karnataka, India

[6]Department of Electronics and Communication Engineering, SRM University-AP, Amaravati 522 240, Andhra Pradesh, India

Corresponding author 1: Viswavardhan Reddy Karna, email: viswavardhank@rvce.edu.in

Corresponding author 2: Neethu S, email: neethus@rvce.edu.in

Corresponding author 3: Vishnu Vardhana Reddy Karna, email: vishnuvardhanareddy.k@srmap.edu.in

Abstract

Accurate district-level wheat-yield forecasting requires models that generalise across locations and future crop years while avoiding spatial and temporal leakage. This study developed an explainable forecasting framework for Madhya Pradesh, India, using 950 harmonised district–crop-year observations from 38 stable districts, ERA5-Land meteorology, historical-yield predictors, and complementary Sentinel-2 indicators. Nineteen model–scenario combinations spanning historical baselines, linear and ensemble models, recurrent networks, and FT-Transformer were screened before a protected final test. Performance was assessed using chronological, rolling-origin, district-grouped, and locked 2020–2022 evaluation, together with SHAP, Integrated Gradients, matched ablation, frozen-model perturbation, sequence-length sensitivity, failure analysis, uncertainty assessment, and satellite–yield associations. SimpleRNN achieved the lowest district-grouped MAE of 304.208 kg ha^{-1}, whereas FT-Transformer performed best under rolling-origin validation with an MAE of 367.610 kg ha^{-1}. Elastic Net achieved the lowest protected-test MAE of 368.498 kg ha^{-1}, outperforming the prior two-year moving-average baseline by 3.79%. Adding historical-yield features reduced LSTM MAE by 50.1% spatially and 59.4% temporally, demonstrating the dominant contribution of

recent productivity history. Prediction-interval coverage deviated from nominal levels by at most 0.877 percentage points, although temporal intervals were wider. High-yield observations remained more difficult to predict. Late-stage NDMI showed the strongest satellite–yield association (Pearson r=0.442; Spearman ρ=0.507). Overall, rigorous validation and strong historical baselines were more influential than model complexity alone.

Keywords: Wheat-yield prediction; leakage-aware validation; explainable machine learning; spatial–temporal generalisation; Sentinel-2; Madhya Pradesh.

1. Introduction

Crop-yield prediction is increasingly important for agricultural planning, food-security assessment, market management, crop insurance, and climate-risk adaptation. Yield is influenced by interacting temporal and spatial factors, including rainfall, temperature stress, solar radiation, soil conditions, irrigation, crop management, and historical productivity. Machine-learning methods are well suited to these heterogeneous relationships because they can represent nonlinear interactions without requiring all relationships to be specified a priori. A systematic review by van Klompenburg et al. identified meteorological, soil, remote-sensing, and historical production variables as major inputs in crop-yield modelling while also highlighting inconsistencies in data preparation, validation, model comparison, and reproducibility [1]. Random Forest models have demonstrated strong capability for regional and global crop-yield prediction using climate and crop-related information [2].

Deep learning provides an alternative for representing complex agricultural relationships. Deep neural networks have successfully modelled interactions among environmental and management variables for crop-yield prediction [3], while remote-sensing information has also been incorporated into deep probabilistic frameworks for regional yield estimation [4]. Machine learning can additionally capture nonlinear yield responses to climate variability, although extrapolation beyond historically observed environmental conditions remains challenging [5]. Hybrid approaches combining machine learning with crop-model information further indicate that data-driven learning and agronomic process knowledge can be complementary rather than competing strategies [6]. Statistical and process-based crop models similarly provide different but complementary perspectives on agricultural responses to environmental change [7].

A central problem, however, is whether apparently accurate models genuinely generalise. Conventional random cross-validation can produce optimistic estimates when observations

have spatial, temporal, or hierarchical dependence [8]. Time-series prediction therefore requires validation procedures that respect temporal ordering [9], and the suitability of a performance-estimation method depends on the forecasting structure and dependence present in the data [10]. These considerations are particularly important for district–crop-year panels because repeated observations from the same districts and neighbouring years are not fully independent.

Recent crop-yield studies further illustrate both the opportunities and limitations of increasingly complex models. Large-scale machine-learning frameworks have demonstrated the feasibility of operational crop forecasting across multiple regions and crops [11], whereas small-data studies show that additional model complexity does not necessarily guarantee substantial forecasting gains [12]. Wheat-yield prediction has benefited from combining satellite and climate information under machine-learning frameworks [13], while antecedent yield, satellite, and climatic information have also proved useful for regional wheat modelling [14]. Hybrid biophysical–machine-learning approaches can improve dynamic wheat forecasting by combining process information with data-driven relationships [15]. Regional European studies similarly demonstrate the potential of machine learning while emphasising the importance of consistent evaluation across heterogeneous production environments [16].

Generalisability across agricultural regions remains an important limitation. District-level prediction using climate reanalysis and remote sensing has shown promise within India [17], while comparisons between panel and time-series approaches indicate that model behaviour can differ substantially across subnational forecasting structures [18]. The present study therefore treats Madhya Pradesh as a defined case-study region rather than assuming that fitted models or performance rankings automatically transfer to Punjab, Haryana, other Indian states, or international wheat systems. Wheat was selected because its Rabi production cycle provides a coherent seasonal forecasting window and because a sufficiently long district-level production record was available for rigorous temporal and spatial evaluation.

Earth observation provides an additional source of information on crop condition. Sentinel-2 imagery has supported high-resolution wheat-yield mapping [19] and within-field wheat-yield assessment [20]. Nevertheless, spatial prediction also requires careful evaluation because neighbouring or environmentally similar observations can inflate apparent predictive skill when separation is inadequate. Spatially structured cross-validation provides a practical mechanism for reducing this risk [21], and empirical evidence shows that model performance can deteriorate substantially when genuinely independent spatial validation is imposed [22].

Interpretability and reliability are equally important for agricultural decision support. Explainable artificial intelligence provides a framework for examining how complex models arrive at their predictions and for identifying limitations in model-dependent explanations [23]. Prediction uncertainty must also be considered because point accuracy alone does not describe the reliability of forecasts under changing environmental conditions; uncertainty-quantification methods for deep learning therefore form an important complementary evaluation dimension [24]. More broadly, uncertainty in neural networks can arise from data limitations, model structure, and distribution shift, requiring explicit reliability assessment rather than assuming that predictive confidence is uniform across observations [25].

Despite these advances, several gaps remain in district-level yield forecasting. Studies often emphasise a single validation design, report only pooled accuracy, or do not clearly separate future-year forecasting from transfer to unseen geographical units. Historical-yield persistence is also not always benchmarked explicitly against recurrent and Transformer-based models, while explainability, perturbation sensitivity, outlier behaviour, uncertainty calibration, and computational requirements are rarely assessed within the same protected evaluation framework.

Accordingly, this study develops an explainable and leakage-aware framework for district-level Rabi wheat-yield prediction in Madhya Pradesh, India, using 950 harmonised district–crop-year observations from 38 stable districts. Historical-yield predictors and ERA5-Land meteorology are evaluated using classical, ensemble, recurrent, and lightweight Transformer models under chronological, rolling-origin, district-grouped, and locked final-test protocols. Complementary Sentinel-2 analysis is used to characterise recent seasonal vegetation and moisture conditions. The study further integrates statistical comparison, explainability, matched ablation, frozen-model perturbation, sequence-length sensitivity, failure analysis, prediction-interval reliability, and computational assessment. Deep models are therefore evaluated as comparative candidates rather than presumed to outperform simpler baselines, and conclusions are restricted to the observed Madhya Pradesh data regime unless independently validated elsewhere.

2. Literature Review

District-level crop-yield forecasting has progressed from conventional statistical methods towards machine-learning, deep-learning, crop-model, and remote-sensing approaches capable of representing nonlinear relationships among historical productivity, weather, and crop condition. Regularised linear models remain important benchmarks because agricultural

predictors are often correlated; Elastic Net combines L_1 and L_2 penalties to provide stable estimation and variable selection under such conditions [26]. Tree-based ensembles provide a complementary nonlinear approach, with Extremely Randomized Trees reducing variance through randomised feature and split selection [27]. For ordered meteorological sequences, LSTM networks provide gated memory mechanisms designed to preserve relevant information across time steps [28]. Attention-based tabular architectures such as FT-Transformer extend deep learning to structured numerical data through feature embeddings and self-attention [29]. Interpretability is increasingly important because predictive accuracy alone does not explain why models succeed or fail. SHAP provides theoretically grounded feature-attribution values for interpreting model predictions [30]. Integrated Gradients provides an attribution mechanism for differentiable neural networks and is particularly suitable for analysing sequential inputs by feature, channel, or time step [31]. Remote sensing also provides complementary indicators of crop condition; SAVI was developed to reduce soil-background effects in vegetation monitoring [32]. MODIS vegetation indices subsequently demonstrated the broader radiometric and biophysical value of vegetation-index monitoring [33]. Moisture-sensitive indices such as NDWI further provide information on vegetation water status that is not fully represented by greenness measures alone [34].

Deep-learning applications have increasingly combined these information sources. LSTM-based integration of meteorological and remotely sensed variables has improved wheat-yield estimation in regional applications [35]. Satellite and weather information have similarly supported machine-learning-based soybean forecasting, demonstrating the value of multi-source predictors beyond a single crop [36]. DeepYield combined convolutional and recurrent representations for crop-yield forecasting, illustrating the potential of architectures that jointly model spatial and temporal information [37]. Probabilistic extensions of deep neural networks have also been developed to quantify uncertainty rather than reporting point predictions alone [38]. Nevertheless, comparisons between Extreme Gradient Boosting and deep-learning approaches show that greater architectural complexity does not automatically produce superior yield estimates [39].

Recent studies have therefore placed increasing emphasis on interpretability and generalisation rather than model complexity alone. Deep crop-yield models have been examined using interpretability methods to identify the variables responsible for their forecasts [40]. Out-of-year Sentinel-2 prediction has shown that temporal transfer is a distinct challenge from within-year fitting and requires explicit evaluation on future seasons [41]. Large-area deep-learning crop models further demonstrate the potential for adaptive yield estimation across spatially

extensive agricultural systems [42]. Recurrent three-dimensional convolutional networks have been used to exploit multi-spectral and multi-temporal imagery simultaneously [43]. High-resolution WorldView-3 and PlanetScope observations have also supported field-scale yield prediction, indicating the value of dense temporal and spatial crop monitoring [44].

Attention mechanisms provide another route for extracting temporally important remotely sensed information. Attention-based deep-learning frameworks have been applied to wheat-yield estimation using remotely sensed indices [45]. UAV hyperspectral imagery combined with machine learning has demonstrated that model performance also depends strongly on spectral information and observation scale [46]. Graph-based deep-learning approaches extend this idea by representing spatial relationships between agricultural observations rather than treating records as independent samples [47]. More recent attention-based wheat models have incorporated explicit uncertainty assessment, reinforcing the need to evaluate both predictive accuracy and reliability [48].

Hybrid modelling has become particularly important because purely data-driven approaches may not capture agronomic structure under distribution shift. Combining crop-model data assimilation with machine learning has improved winter-wheat forecasting by integrating physiological information with learned relationships [49]. In-season prediction studies also show that the relative contribution of climate forecasts and satellite observations changes across crop-development phases [50]. The synergistic use of SAR, optical imagery, and meteorological data similarly improves yield prediction by combining complementary information sources [51]. Assimilation of observations into WOFOST demonstrates how process-based crop models can provide an agronomic framework for regional wheat-yield estimation [52].

Water and vegetation dynamics remain important sources of within-season information. Including crop water deficit can strengthen remotely sensed yield models by representing stress conditions that vegetation imagery alone may not fully capture [53]. CubeSat observations assimilated into crop models have enabled early-season prediction of within-field yield variability [54]. Sentinel-2 leaf-area information has likewise been coupled with crop-growth modelling to predict spatial yield variation [55]. At broader scales, photosynthesis and meteorological stress indices have supported nationwide crop-yield estimation [56]. Wheat studies combining spectral vegetation indices with meteorological variables further indicate that remote-sensing signals and weather information provide complementary predictive information [57].

Recent work has consequently moved towards flexible multi-source frameworks. Versatile crop-yield estimators have been developed to accommodate heterogeneous environmental and observational information [58]. Hybridisation of process-based models, remote sensing, and machine learning has been proposed for spatial prediction of wheat yield and quality [59]. Deep-learning models integrating multiple environmental data sources further illustrate the continuing expansion of wheat-yield forecasting beyond single-source predictors [60].

Explainability and reliability remain essential when these models are intended for agricultural decision support. Tree-specific SHAP methods enable local explanations to be aggregated into global understanding of fitted model behaviour [61]. However, dependence among predictors can affect Shapley-value interpretation, requiring caution when correlated agricultural variables are analysed [62]. This concern supports retaining inherently interpretable models as benchmarks rather than assuming that post hoc explanations fully resolve black-box behaviour [63]. Distribution-free predictive inference provides a principled basis for constructing regression uncertainty intervals without relying on strong parametric assumptions [64]. Jackknife+ methods further extend predictive inference by providing finite-sample interval guarantees under appropriate conditions [65]. Proper scoring rules are therefore useful for evaluating uncertainty estimates jointly in terms of calibration and sharpness rather than coverage alone [66].

Overall, the literature supports a comparative framework rather than a presumption that one model family is universally superior. District-level crop-yield forecasting should therefore combine transparent baselines, nonlinear models, sequential architectures, remote-sensing information, leakage-aware temporal and spatial validation, interpretable prediction mechanisms, and explicit uncertainty assessment. This provides the basis for evaluating whether additional model complexity produces genuine and transferable gains rather than improvements confined to a particular validation design.

3. Methodology

3.1 Research Design, Study Scope, and Prediction Objectives

3.1.1 Overall Research Design

This study used a retrospective, secondary-data predictive-modelling design for district-level Rabi wheat-yield estimation in Madhya Pradesh, India. The analytical unit was a stable district–crop-year observation, with official wheat yield (kg ha^{-1}) as the response. Agricultural statistics, administrative boundaries and historical district crosswalks, ERA5-Land

meteorology, historical-yield predictors, and complementary Sentinel-2 data were integrated using Python and Google Earth Engine.

The framework comprised data harmonisation, target reconstruction, leakage-controlled feature engineering, model development, temporal and spatial validation, one-time locked final testing, explainability and robustness analysis, prediction-reliability assessment, computational evaluation, and complementary satellite–yield analysis. All feature definitions, preprocessing rules, validation procedures, model-selection decisions, and canonical reporting models were established before accessing the locked final-test period.

Figure 1 summarises the complete leakage-aware workflow from data acquisition and stable-district harmonisation to model evaluation, explainability, robustness, uncertainty assessment, computational reporting, and Sentinel-2 analysis.

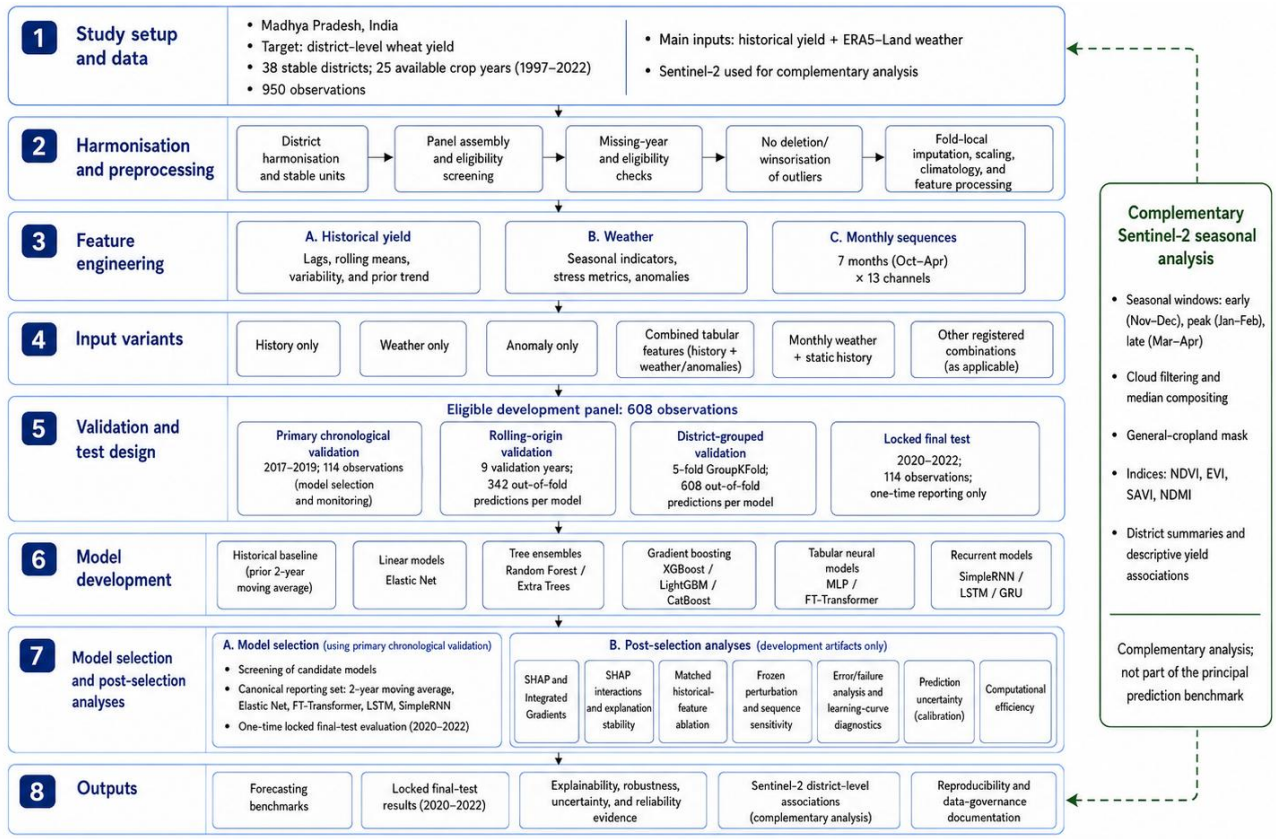


Figure 1. Leakage-aware methodological workflow for district-level Rabi wheat-yield forecasting and complementary Sentinel-2 seasonal association analysis in Madhya Pradesh, India.

3.1.2 Study Area and Madhya Pradesh Selection

Madhya Pradesh was selected because it provides a long, consistent district-level wheat-production record, reconstructable historical administrative boundaries, complete long-term meteorological coverage, and adequate recent Sentinel-2 observations. Its variation in wheat productivity, cultivated area, agroclimatic conditions, irrigation availability, and agricultural

intensity also supports evaluation of both future-year forecasting and transfer to districts excluded from model fitting.

The study is a state-specific empirical assessment rather than evidence of universal geographical generalisability. Although the framework may be adaptable elsewhere, the reported performance has not been externally validated for Punjab, Haryana, Maharashtra, other Indian states, or wheat-growing regions outside India. Application to these regions would require region-specific data harmonisation, model retraining, and independent temporal and spatial validation.

3.1.3 Wheat-Crop Selection and Temporal Framework

Wheat was selected because it is an important Rabi crop with a well-defined seasonal cycle and a sufficiently long official district-level production record. Its crop calendar allowed consistent representation of meteorological conditions from October of the sowing year to April of the following year, while the historical record supported comparison of recent productivity, yield persistence, variability, and trend with meteorological predictors.

Crop years were indexed by their starting year; for example, crop-year start 2019 represented the 2019–2020 Rabi season. The meteorological season extended from 1 October of the starting year to 30 April of the following year, with 1 May used as the exclusive date-filtering endpoint. Agricultural and meteorological records covered 1997–1998 to 2022–2023, except for the unavailable 2009–2010 agricultural record. Sentinel-2 analysis covered 2017–2018 to 2022–2023 because consistent harmonised Level-2A surface-reflectance data were available from 2017 onward. Soybean, rice, maize, and other crops were outside the study scope because they would require crop-specific seasonal windows, feature definitions, model development, and external validation.

3.1.4 Sample-Size Rationale and Effective Independence

The harmonised archive contained 950 district–crop-year observations from 38 stable districts and 25 available crop years. Because districts were repeatedly observed across years, these records were not fully independent; spatial persistence and temporal dependence reduced the effective information available for fitting highly parameterised models.

Historical-feature eligibility and separation of the locked final-test period produced a common development panel of 608 observations, with a further 114 observations reserved for the locked 2020–2022 test. These constraints motivated compact neural architectures, constrained development-stage searches, strong persistence and regularised linear baselines, early stopping,

weight regularisation, gradient clipping, fixed random seeds, and leakage-aware spatial and temporal validation.

Deep-learning models were therefore treated as controlled comparative candidates rather than presumed superior. Conclusions were restricted to the observed district-level panel, evaluated model configurations, and specified validation protocols.

3.1.5 Prediction Tasks and Generalisation Targets

Four complementary evaluation tasks were defined. Primary chronological validation used later development years to screen model–scenario candidates and establish the canonical reporting set. Rolling-origin validation assessed future-year forecasting using only observations preceding each validation crop year. District-grouped validation assessed spatial transfer by withholding complete stable districts from model fitting, while the locked 2020–2022 final test provided a one-time evaluation of the frozen canonical models.

These protocols targeted different forms of generalisation. District-grouped validation measured transfer to untrained geographical units for which historical predictors remained available, whereas rolling-origin validation assessed forecasting under temporal distribution shift. Performance under one protocol was therefore not assumed to imply equivalent performance under another.

3.1.6 Locked Final-Test Policy

Crop-year starts 2020, 2021, and 2022 were reserved as the locked final-test period before model development. These observations were excluded from feature selection, preprocessing decisions, climatological-reference estimation, architecture and hyperparameter selection, candidate promotion, statistical-procedure selection, and prediction-interval calibration.

The final test was executed once after the feature registry, preprocessing rules, validation assignments, statistical plan, random seeds, model configurations, and canonical reporting set had been frozen. Models excluded during development were not reconsidered after final-test inspection.

Subsequent explainability, perturbation, sequence-sensitivity, reliability, computational, and reporting analyses used saved development-stage predictions, fitted pipelines, neural checkpoints, and development-only observations where applicable. Locked final-test results were retained solely for final reporting and were not reused for model selection or analytical refinement.

3.2 Data Sources, Selection Rationale, and Missing Covariates

The study integrated official agricultural statistics, administrative-boundary data, meteorological reanalysis, surface-reflectance imagery, optical-quality information, and land-cover products. Table 1 summarises the source, coverage, extracted information, and analytical role of each dataset. Agricultural, boundary, and ERA5-Land data supported the principal predictive analysis, whereas Sentinel-2 and land-cover products supported a complementary district-level vegetation and moisture assessment.

Table 1. Data sources and their analytical roles

Data component	Source	Coverage and resolution	Information used	Analytical role
Agricultural statistics	Directorate of Economics and Statistics, Government of India [67]	District–crop-year; 1997–1998 to 2022–2023, excluding 2009–2010	Cultivated area, production, reported yield, district, crop, season, and crop year	Target reconstruction and historical-yield feature generation
Administrative boundaries	Survey of India Administrative Boundary Data Base [68]	Contemporary district polygons	District names, Local Government Directory codes, state identifiers, and geometry	Reconstruction of stable historical district units
Meteorological reanalysis	ERA5-Land Daily Aggregated [69]	Daily; approximately 11.1 km; 1997–1998 to 2022–2023	Precipitation, air and dew-point temperature, radiation, and wind	Seasonal, monthly, anomaly, event, and sequential predictors
Surface reflectance	Harmonised Sentinel-2 Level-2A	10–20 m; 2017–2018 to 2022–2023	Blue, red, near-infrared, and short-wave-	Complementary vegetation and

	Surface Reflectance [70]		infrared reflectance	moisture assessment
Optical-quality information	Cloud Score+ S2 Harmonized [71]	Sentinel-2 pixel grid	cs_cdf optical-quality score	Cloud and atmospheric-quality filtering
Dynamic land cover	Dynamic World V1 [72]	10 m	Crop-class probability	Supporting land-cover assessment
Static land cover	ESA WorldCover 2021 v200 [73]	10 m	Class 40 cropland	Fixed general-cropland mask

3.2.1 Agricultural Statistics

District-level agricultural statistics were obtained from the Directorate of Economics and Statistics through the Open Government Data Platform India [67] and filtered for Madhya Pradesh, wheat, and the Rabi season.

District names were standardised, crop-year labels separated into starting and ending years, and cultivated area, production, and reported yield converted to numeric form; conversion failures were retained as missing during quality assessment. The original extract contained 1,217 records, 52 agricultural-statistics-compatible district names, and 25 available crop years. Wheat yield was independently reconstructed from cultivated area and production as:

$$Y_{d,t} = \frac{1000 P_{d,t}}{A_{d,t}}, \quad (1)$$

where $Y_{d,t}$ is the reconstructed wheat yield for district d and crop year t, expressed in kg ha $^{-1}$; $P_{d,t}$ is wheat production in tonnes; $A_{d,t}$ is cultivated wheat area in hectares; and the factor 1000 converts tonnes to kilograms.

Reconstruction was necessary because area and production are additive when administrative units are merged, whereas district yields cannot be validly combined by an unweighted mean. Reconstructed and reported yields were compared to detect unit or data-entry inconsistencies. Duplicate district–crop-year keys, invalid area or production values, and non-computable yields were checked before spatial harmonisation. The unavailable 2009–2010 agricultural record remained missing and was neither reconstructed nor imputed.

3.2.2 Administrative Boundaries and Stable-District Reconstruction

District polygons were obtained from the Survey of India Administrative Boundary Data Base [68] and filtered to Madhya Pradesh, yielding 55 contemporary districts. Retained attributes included district name, Local Government Directory code, state identifiers, and geometry.

The boundary layer was checked for duplicate identifiers, missing or invalid geometries, unexpected district names, and coordinate-reference issues. Invalid geometries were repaired, and the final layer was transformed to EPSG:4326.

Because several present-day districts were created during the study period, the contemporary boundaries could not be joined directly to the full agricultural record. A documented historical crosswalk was therefore used to assign recent child districts to consistent parent units and dissolve the 55 polygons into 38 stable analytical districts. This preserved consistent district definitions through time and prevented administrative changes from being interpreted as productivity changes. The detailed crosswalk and aggregation procedure are described in Section 3.3.

3.2.3 ERA5-Land Meteorological Data

Daily meteorological data were obtained from the ECMWF/ERA5_LAND/DAILY_AGGR collection [69]. ERA5-Land was selected for its spatial completeness, temporal continuity, and consistent variable definitions across all stable districts and crop years.

Extracted variables included precipitation, minimum, maximum, and mean 2-m air temperature, 2-m dew-point temperature, downward surface solar radiation, and 10-m zonal and meridional wind components. Temperature was converted from Kelvin to °C, precipitation from metres to millimetres, and radiation from J m^{-2} to MJ m^{-2}; small negative precipitation artefacts were set to zero.

Wind speed and relative humidity were derived, and daily temperature range, growing-degree days, rainy and heavy-rainfall days, heat-stress exposure, maximum one-day rainfall, and maximum dry-spell duration were calculated. A 5 °C base temperature was used for wheat growing-degree days. Meteorological fields were aggregated over stable district polygons at approximately 11,132 m and organised into October–April seasonal summaries and monthly sequences. Training-partition-specific anomaly construction is described in Section 3.4.

ERA5-Land was not directly benchmarked against CHIRPS, NASA POWER, India Meteorological Department products, or MERRA; the results therefore reflect a consistently harmonised ERA5-Land-based framework rather than a comparison of meteorological products.

3.2.4 Sentinel-2 and Land-Cover Data

Sentinel-2 Level-2A surface reflectance was obtained from the COPERNICUS/S2_SR_HARMONIZED collection [70], selected for numerical consistency across processing baselines, suitable visible, near-infrared, and short-wave-infrared bands, and adequate recent coverage. Reflectance values were multiplied by 0.0001 before spectral-index calculation.

Cloud Score+ data from GOOGLE/CLOUD_SCORE_PLUS/V1/S2_HARMONIZED [71] were linked to Sentinel-2 scenes using common image identifiers and combined with Sentinel-2 quality information to exclude cloud-contaminated and unsuitable pixels.

Dynamic World V1 crop probability was examined as supporting land-cover information [72], while final publication statistics used ESA WorldCover v200 class 40 as a fixed general-cropland mask [73]. This mask was not treated as a wheat-specific classification; therefore, Sentinel-2 results represent district-level spectral conditions over mapped general cropland associated with official wheat statistics rather than field-level wheat observations.

Sentinel-2 was not benchmarked against MODIS, Landsat, or Planet imagery. The satellite component was intended as a complementary seasonal association analysis using one consistently processed image source, not as a comparison of remote-sensing platforms.

3.2.5 Unavailable Soil, Terrain, Management, and Economic Variables

Consistent district–crop-year data were unavailable for soil texture, soil organic carbon, soil pH, elevation, slope, irrigation coverage, groundwater availability, fertiliser application, cropping intensity, sowing date, market price, and other management inputs. These variables were therefore excluded, and no indirect proxy was treated as an equivalent direct measurement.

Historical-yield predictors may partly reflect persistent local conditions related to soil, irrigation, management, technology, and reporting, but they were used only as predictive indicators and not as causal measures. The absence of direct soil, irrigation, and management variables therefore limits agronomic attribution and is addressed in the Discussion and study limitations.

3.3 Spatial and Temporal Data Harmonisation

3.3.1 Stable-District Construction

Several present-day districts were created by subdividing earlier administrative units, making the 55 contemporary polygons unsuitable for direct use across the full agricultural record. They

were therefore reconstructed into 38 stable analytical districts to maintain consistent spatial definitions and avoid artificial changes in cultivated area, production, and yield.

The crosswalk used standardised district names, Local Government Directory identifiers, and documented parent–child relationships. Mappings combined Pandhurna–Chhindwara, Ashoknagar–Guna, Harda–Hoshangabad, Katni–Jabalpur, Alirajpur–Jhabua, Burhanpur–Khandwa, Barwani–Khargone, Dindori–Mandla, Neemuch–Mandsaur, Sheopur–Morena, Mauganj–Rewa, Maihar–Satna, Anuppur and Umaria–Shahdol, Agar Malwa–Shajapur, Singrauli–Sidhi, and Niwari–Tikamgarh. Narmadapuram was aligned with Hoshangabad and Narsimhapur with Narsinghpur.

An intermediate 52-unit geography matched the agricultural source before dissolution into 38 stable districts. Geometries were checked, repaired, transformed to EPSG:4326, and audited to ensure each source unit mapped to one stable district.

3.3.2 Agricultural Aggregation and Yield Validation

For each stable district and crop year, cultivated area and production were summed across contributing source districts using the administrative crosswalk, and wheat yield was recalculated from the aggregated values using Equation (1).

Source-district yields were not averaged directly, as this would ignore differences in cultivated area. Recalculation after aggregation preserved appropriate area weighting and consistency with the stable spatial units.

Reconstructed yields were checked against source information for unit, aggregation, or transcription inconsistencies. Harmonisation produced 950 stable district–crop-year observations across 38 districts and 25 available crop years, with no duplicate keys, unmatched units, invalid area or production values, or unresolved target inconsistencies.

3.3.3 Crop-Year Definition and Seasonal Alignment

Crop years were indexed by their starting year; for example, crop-year start 2019 represented the 2019–2020 Rabi season. Agricultural targets were aligned with meteorological observations from 1 October of the starting year to 30 April of the following year, with 1 May as the exclusive date-filtering endpoint.

This alignment ensured that seasonal and monthly predictors corresponded to the correct wheat-production record. Each recurrent sequence contained seven monthly positions in the fixed October–April order.

3.3.4 Missing-Year Policy and Analytical Eligibility

The complete 2009–2010 agricultural record was unavailable and was not reconstructed or imputed. As summarised in Table 2, crop years whose historical-feature windows intersected this missing record were excluded from matched modelling. Crop-year starts 1997–1999 were used only for historical-feature initialisation; 2000–2008 and 2013–2019 formed the eligible development period; 2010–2012 remained in the archive but were excluded from common modelling; and 2020–2022 were reserved as the locked final test.

Table 2. Crop-year coverage and analytical treatment

Crop-year starts	Stable observations	Analytical treatment	Rationale
1997–1999	114	Historical-feature initialisation only	Complete preceding historical windows were unavailable
2000–2008	342	Eligible development period	Consecutive historical windows were available
2009	0	Unavailable	The complete 2009–2010 agricultural record was absent
2010–2012	114	Retained in archive but excluded from matched modelling	Required historical windows intersected the unavailable crop year
2013–2016	152	Eligible development period	Consecutive historical observations were again available

2017–2019	114	Primary chronological validation within development	Later development years used for candidate screening
2020–2022	114	Locked final test	Withheld before model selection
2017–2022	228	Complementary Sentinel-2 assessment	Consistent Level-2A coverage was available

The common development panel contained 608 observations from 38 districts across 16 eligible crop years. The 228-observation Sentinel-2 panel overlaps the development and final-test periods and was analysed separately.

3.3.5 Dataset Alignment and Quality Assurance

All datasets were aligned using a composite stable-district and crop-year-start key linking agricultural targets, historical-yield features, seasonal and monthly weather data, satellite statistics, validation assignments, predictions, uncertainty estimates, perturbation records, and explanation outputs.

Tabular and sequential datasets were audited for unmatched identifiers, duplicate district–crop-year keys, target or crop-year disagreement, missing or non-finite values, invalid geometries, incomplete or misordered October–April sequences, ineligible historical-feature windows, and chronological inconsistencies.

Each accepted recurrent input contained all seven monthly positions in the prescribed order. District identity was retained for alignment, mapping, grouped validation, reliability assessment, and residual reporting, but excluded from the primary predictor matrices to prevent direct district memorisation.

3.3.6 Authoritative Sample Accounting

Table 3 presents the definitive sample counts used in the study. Because the evaluation sets overlap, these counts should not be summed to derive a total sample size.

Table 3. Authoritative data and evaluation sample accounting

Component	Count	Interpretation
Original agricultural records	1,217	Records before stable-district harmonisation

Agricultural-source district names	52	District names compatible with the agricultural dataset
Contemporary boundary polygons	55	Present-day Madhya Pradesh districts
Stable analytical districts	38	Historically consistent spatial units
Available crop years	25	All years except the unavailable 2009–2010 record
Harmonised district–crop-year observations	950	districts × available crop years
Common development observations	608	districts × eligible crop years
Primary chronological-validation observations	114	2017–2019; subset of the development panel
District-grouped out-of-fold predictions per model	608	One prediction for each development observation
Rolling-origin out-of-fold predictions per model	342	Nine held-out crop years × 38 districts
Locked final-test observations	114	2020–2022; three crop years × 38 districts
Sentinel-2 district–crop-year observations	228	2017–2022; six crop years × 38 districts

3.4 Feature Engineering and Preprocessing

The predictor framework separated prior productivity, absolute meteorological conditions, climate anomalies, and ordered monthly weather information. This separation enabled the contribution of each information source to be evaluated under matched validation and preprocessing conditions. Satellite-derived variables were retained in a separate recent-period panel and were not included in the primary predictive feature matrices. The predictor groups, representations, and leakage controls are summarised in Table 4.

Table 4. Predictor groups, representations, and preprocessing controls

Predictor group	Principal variables	Representation	Leakage control
Historical yield	One-year lag, two- and three-year means, two- and three-year standard	District–crop-year tabular input	Only crop years preceding the target year were used

	deviations, coefficients of variation, and prior trend		
Seasonal weather	Rainfall, temperature, humidity, radiation, wind, growing-degree days, rainfall events, heat exposure, and dry spells	District–season tabular input	Transformations were estimated within each training partition
Climate anomalies	Absolute and standardised rainfall and temperature departures	District–season tabular input	Climatological references were estimated from training observations only
Monthly weather sequence	Seven October–April positions with 13 meteorological channels	tensor	Monthly order was fixed, and preprocessing used training observations only
Static recurrent input	Eight historical-yield features	Separate static branch	Imputation and scaling were fitted within each training partition
Satellite variables	NDVI, EVI, SAVI, NDMI, seasonal-stage summaries, and stage changes	Separate recent-period district–crop-year panel	No satellite values were reconstructed before 2017

3.4.1 Historical-Yield Features

Historical-yield features were generated chronologically for each stable district. The frozen eight-feature block represented immediate persistence, recent productivity, short-term and relative variability, and prior yield trend.

The one-year lag was defined as

$$L_{d,t}^{(1)} = Y_{d,t-1}, \tag{2}$$

where $L_{d,t}^{(1)}$ is the lagged yield for district d and target crop year t, and $Y_{d,t-1}$ is the observed yield in the immediately preceding crop year.

Mean yield over the preceding K crop years was calculated as

$$\bar{Y}_{d,t}^{(K)} = \frac{1}{K}\sum_{k=1}^{K} Y_{d,t-k}, K \in \{2,3\}, \quad (3)$$

where K denotes the historical window. Two- and three-year means represented recent productivity while smoothing isolated annual fluctuations.

Historical variability was calculated as

$$s_{d,t}^{(K)} = \sqrt{\frac{\sum_{k=1}^{K}\left(Y_{d,t-k} - \bar{Y}_{d,t}^{(K)}\right)^2}{K-1}}, K \in \{2,3\}, \quad (4)$$

where $s_{d,t}^{(K)}$ is the sample standard deviation of prior yield. Relative variability was represented by the coefficient of variation:

$$CV_{d,t}^{(K)} = \frac{s_{d,t}^{(K)}}{\bar{Y}_{d,t}^{(K)}}, K \in \{2,3\}. \quad (5)$$

This dimensionless measure enabled variability comparison across districts with different productivity levels.

The eighth feature was a prior trend term derived exclusively from observations preceding the target year, indicating whether recent district yield had increased, decreased, or remained stable. Its calculation and window were fixed before validation.

Exact crop-year matching was required. An earlier observation was not substituted when an intervening crop year was missing, and current or future yields were excluded from all historical calculations. These predictors represented prior productivity and were not interpreted as causal measures of soil, irrigation, management, technology, or other unobserved agronomic factors.

3.4.2 Seasonal Weather and Stress Features

Daily ERA5-Land observations were aggregated over the October–April Rabi wheat season. Seasonal predictors included total rainfall, maximum one-day rainfall, rainy- and heavy-rainfall frequency, maximum dry-spell duration, minimum, maximum and mean temperature, temperature range, relative humidity, solar radiation, wind speed, growing-degree days, and heat-stress-day frequency.

Rainy and heavy-rainfall days were defined using precipitation thresholds of ≥1 mm and ≥10 mm, respectively. Dry days used the <1 mm threshold, with maximum dry-spell duration representing the longest consecutive dry period. Heat-stress days were defined by maximum temperature >35 °C, while growing-degree days used a 5 °C base temperature.

Rainfall- and temperature-event indicators were calculated at the ERA5-Land pixel level and then averaged within each stable district, allowing fractional district-level counts. Maximum dry-spell duration and maximum one-day rainfall were similarly summarised as district means of pixel-level seasonal values.

3.4.3 Weather-Anomaly Features

Absolute meteorological variables represented observed seasonal conditions, while anomaly features measured departures from training-based climatological references. For meteorological variable j, the absolute anomaly was defined as

$$A_{d,t,j} = X_{d,t,j} - \mu_j, \tag{6}$$

where $A_{d,t,j}$ is the anomaly for district d, crop year t, and variable j; $X_{d,t,j}$ is the observed seasonal value; and μ_j is the climatological mean estimated only from the applicable training partition. Positive and negative values indicate conditions above and below the training reference, respectively.

Standardised anomalies were calculated as

$$Z_{d,t,j} = \frac{A_{d,t,j}}{\sigma_j}, \tag{7}$$

where σ_j is the training-partition standard deviation, expressing departures in units of historical variability.

During rolling-origin validation, μ_j and σ_j were estimated only from preceding crop years. District-grouped validation excluded held-out districts from climatological estimation, using the predefined pooled training-only reference where district-specific values were unavailable. Locked final-test observations were excluded from climatological estimation and anomaly construction, preventing leakage from future years or held-out districts.

3.4.4 Monthly Weather Sequences

Each eligible district-crop-year observation comprised seven ordered monthly positions from October to April, each containing 13 meteorological channels: rainfall total, mean pixel-level maximum one-day rainfall, rainy-day frequency, heavy-rainfall frequency, mean minimum temperature, mean maximum temperature, mean temperature, mean temperature range, mean relative humidity, total solar radiation, mean wind speed, growing-degree-day accumulation, and heat-stress-day frequency.

The recurrent weather input was represented as

$$\mathrm{X} \in \mathbb{R}^{N\times 7\times 13}, \tag{8}$$

where N is the number of eligible district-crop-year observations, with seven monthly positions and 13 meteorological channels, preserving intra-seasonal weather ordering.

A separate static input matrix was defined as

$$\mathbf{S} \in \mathbb{R}^{N \times 8}, \tag{9}$$

where the eight columns corresponded to the historical-yield features described in Section 3.4.1, enabling recurrent models to combine monthly weather dynamics with prior district productivity.

Sequences were accepted only when all seven months were present exactly once, correctly ordered, and matched the corresponding stable district, crop year, and response. Imputation and scaling parameters were estimated only from the training partition and applied unchanged to validation and test sequences.

3.4.5 Feature Scenarios

Seven tabular feature scenarios were predefined: history-only, weather-only, anomaly-only, weather-plus-anomaly, history-plus-weather, history-plus-anomaly, and fully combined. These respectively represented the eight historical-yield features, absolute seasonal meteorology, rainfall and temperature anomalies, and their specified combinations.

The scenarios tested the predictive value of prior district productivity, meteorological conditions, climate departures, and whether weather information added value beyond historical persistence.

Recurrent experiments compared monthly weather sequences alone with sequences supplemented by the static historical-yield branch. Architecture, folds, preprocessing, optimisation, and random seeds were held constant to isolate the contribution of prior district productivity.

3.4.6 Fold-Local Preprocessing

All data-dependent preprocessing was estimated independently within each training partition, including predictor imputation, feature scaling, climatological-reference estimation, anomaly construction, feature selection, and neural-network target standardisation. Fitted transformations were applied unchanged to the corresponding validation or test data.

For continuous predictor x_j,

$$x_{i,j}^{*} = \frac{x_{i,j} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}}}, \tag{10}$$

where $x_{i,j}$ is the original value and μ_j^{train} and σ_j^{train} are the training-partition mean and standard deviation. This placed predictors on comparable scales without allowing validation or test data to influence preprocessing.

Missing responses were not imputed, and validation or locked final-test outcomes were excluded from imputation, scaling, feature inclusion, climate-reference estimation, and model configuration.

For neural models, the response was standardised as

$$y_i^* = \frac{y_i - \mu_y^{\text{train}}}{\sigma_y^{\text{train}}}, \tag{11}$$

where μ_y^{train} and σ_y^{train} were estimated from the training partition. Predictions were returned to the original scale using

$$\hat{y}_i = \hat{y}_i^* \sigma_y^{\text{train}} + \mu_y^{\text{train}}, \tag{12}$$

where $\hat{y}_i^*$ is the standardised neural output and $\hat{y}_i$ is the final prediction in kgha^{-1}. All performance metrics were calculated after inverse transformation.

3.5 Models, Comparator Selection, and Training

The benchmark compared historical baselines with linear, bagging, boosting, feed-forward, recurrent, and attention-based models to assess whether added complexity improved upon yield persistence and recent district productivity. Given the moderate effective sample size and spatial–temporal dependence, compact neural architectures and constrained development-stage searches were used. Table 5 summarises the evaluated model groups, input structures, and principal training controls.

Table 5. Model groups, input structures, and principal training controls

Model group	Evaluated models	Input structure and analytical role	Principal controls
Historical baselines	Overall training mean, district historical mean, previous-year persistence, prior two-year moving average, and prior three-year moving average	Historical target information used to establish minimum operational performance	Recalculated within each training partition using only eligible prior observations
Linear models	Ordinary least squares, Ridge, and Elastic Net	Predefined tabular feature scenarios	Training-partition imputation and

			scaling; regularisation selected during development
Bagging ensembles	Random Forest and Extra Trees	Tabular predictors with nonlinear effects and interactions	300 trees; candidate depth, minimum leaf size, and feature-fraction settings
Boosting models	HistGradientBoosting, XGBoost, LightGBM, and CatBoost	Sequentially fitted nonlinear tabular learners	Compact development-stage parameter searches
Feed-forward model	Multilayer perceptron	Scaled tabular predictors	Dropout, weight decay, early stopping, and best-checkpoint restoration
Recurrent models	SimpleRNN, LSTM, and GRU	Seven monthly positions and 13 weather channels, with a separate historical branch where applicable	Fixed seeds, gradient clipping, learning-rate scheduling, early stopping, and checkpoint restoration
Attention model	FT-Transformer	Numerical tabular predictors	Numerical tokenisation, compact self-attention blocks, dropout, and feed-forward transformations
Frozen reporting set	Prior two-year moving average, Elastic Net, FT-Transformer, LSTM, and SimpleRNN	Principal cross-protocol and locked-test comparison	Selected and frozen before final-test access

3.5.1 Historical Baselines

Five historical baselines were evaluated before the machine-learning models: the overall training mean, district historical mean, previous-year persistence, and prior two- and three-year

moving averages. These represented non-spatial average yield, longer-term district productivity, immediate persistence, and smoothed recent productivity.

All baseline predictions were recalculated within each training partition using only information available before the prediction year. Exact crop-year continuity was required, and earlier observations were not substituted for missing intervening years. These baselines tested whether more complex models provided meaningful gains over historical information available at the same forecast horizon. The prior two-year moving average was retained as the canonical historical baseline for cross-protocol comparisons.

3.5.2 Classical Linear and Ensemble Models

Ordinary least squares provided an unregularised linear reference, while Ridge regression used L_2 regularisation to stabilise coefficients under predictor correlation. Elastic Net combined $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ penalties:

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right) \right], \tag{13}$$

where $X \in \mathbb{R}^{n \times p}$ is the predictor matrix, y the observed-yield vector, and β the coefficient vector. $\lambda \geq 0$ controls regularisation strength and $\alpha \in [0,1]$ balances the L_1 and L_2 penalties. Elastic Net was suitable for the correlated historical and meteorological predictors and moderate effective sample size, providing coefficient shrinkage and improved interpretability.

Random Forest and Extra Trees represented bagging-based nonlinear models, with ensemble prediction

$$\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x), \tag{14}$$

where B is the number of trees and $T_b(x)$ is the prediction from tree b. Random Forest used bootstrap sampling and random feature selection, whereas Extra Trees additionally randomised split thresholds.

HistGradientBoosting, XGBoost, LightGBM, and CatBoost represented boosting methods in which successive learners corrected preceding errors:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} \ell \left[y_i, \hat{y}_i^{(m-1)} + f_m(x_i) \right] + \Omega(f_m), \tag{15}$$

where m is the boosting stage, f_m the newly added weak learner, $\ell(\cdot)$ the prediction loss, and $\Omega(f_m)$ the complexity penalty. The implementations differed in tree construction, regularisation, histogram approximation, and feature handling, and were evaluated using compact development-stage parameter searches.

3.5.3 Feed-Forward and Recurrent Neural Networks

A compact multilayer perceptron was evaluated as a feed-forward comparator for scaled tabular inputs. It applied nonlinear hidden-layer transformations to produce a continuous yield prediction but did not preserve monthly ordering.

SimpleRNN, LSTM, and GRU processed the seven-month October-April weather sequence. Weather-only configurations used the final sequential representation directly, whereas combined configurations concatenated it with a separately encoded branch containing the eight historical-yield features before regression.

For SimpleRNN, the hidden state was

$$h_t = \phi(W_x x_t + W_h h_{t-1} + b_h), \tag{16}$$

with final prediction

$$\hat{y} = W_y h_T + b_y, \tag{17}$$

where x_t is the meteorological input at month t, h_t the hidden state, W_x, W_h, W_y learned weights, b_h, b_y biases, and $T = 7$. The hidden state summarises information accumulated across the ordered sequence.

The LSTM controlled information flow through input, forget, and output gates:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{18a}$$
$$f_t = \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big), \tag{18b}$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{18c}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{18d}$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{18e}$$
$$h_t = o_t \odot \tanh(c_t). \tag{18f}$$

Here, i_t, f_t, o_t are the gates, $\tilde{c}_t$ the candidate state, c_t the memory state, W and U learned weight matrices, and b learned biases; $\odot$ denotes element-wise multiplication.

The GRU used update and reset gates as a more compact gated alternative to LSTM. SimpleRNN tested whether basic recurrence was sufficient for the short sequence, while LSTM and GRU assessed the value of gating. The static branch allowed historical productivity to be combined with ordered weather information.

3.5.4 FT-Transformer and Scope of the Deep-Learning Benchmark

The FT-Transformer was evaluated as a compact attention-based model for numerical tabular predictors. Each standardised numerical feature x_j was converted into a learned token representation:

$$\mathrm{z}_j = x_j \mathrm{e}_j + \mathrm{b}_j, \tag{19}$$

where e_j and b_j are the learned feature-specific embedding and bias, respectively. Self-attention was computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{20}$$

where Q, K, and V are learned query, key, and value representations and d_k is the key dimension. Attention blocks were followed by feed-forward transformations, dropout, and a regression head.

FT-Transformer was selected before final-test access as a representative compact tabular attention model for numerical predictors and nonlinear feature interactions, rather than as a benchmark of all modern deep-learning architectures.

TabNet, SAINT, NODE, TabTransformer, and TabPFN were not evaluated. Temporal Fusion Transformer, Informer, Autoformer, PatchTST, TimesNet, N-HiTS, and Chronos were also excluded because the sequential benchmark contained only seven monthly positions and prioritised compact recurrent models. These omissions limit architecture-level generalisation but do not imply inferior performance of the excluded methods. Adding architectures after locked final-test inspection would introduce post hoc selection risk; their evaluation is therefore reserved for future externally validated benchmarking.

3.5.5 Neural Training and Reproducibility Controls

Neural models used the standardised response from Section 3.4.6, AdamW optimisation, and mean-squared error for training. Validation MAE on the original yield scale controlled early stopping and best-checkpoint selection.

All canonical neural models used batch size 64, a maximum of 250 epochs, early-stopping patience of 30, gradient clipping at 5.0, and ReduceLROnPlateau with factor 0.50 after 10 epochs without improvement. The checkpoint with the lowest validation MAE was restored. Model-specific settings are summarised in Table 6.

Table 6. Frozen architecture and training configurations of the canonical neural models

Model	Input structure	Hidden or token dimension	Layers or blocks	Additional architecture settings	Dropout	Batch size	Learning rate	Weight decay	Maximum epochs	Early-stopping patience	Frozen final epochs	Trainable

FT-Transformer	8 numerical tabular features	32	1 Transformer block	4 attention heads; feed-forward	0.10	64	0.0010	0.0001	250	30	27	9,185
LSTM	13 sequential weather inputs and 8 static	16	1 recurrent layer	Unidirectional recurrence	0.10	64	0.0010	0.0001	250	30	49	2,505
SimpleRNN	13 sequential weather inputs and 8 static	64	2 recurrent layers	Unidirectional recurrence	0.30	64	0.0005	0.0010	250	30	51	16,929

Frozen final epochs were determined from development-stage best-epoch records before locked final-test access. All 42 records were available for LSTM and SimpleRNN, giving 49 and 51 epochs, respectively. For FT-Transformer, 41 of 42 records were available; sensitivity analysis showed that the ceiling-median could only be 26 or 27, so 27 epochs were conservatively frozen without imputing the missing record.

Three fixed seeds—20260721, 20260722, and 20260723—were used for eligible neural runs. Seed-level predictions were retained separately and averaged arithmetically; no seed was selected retrospectively.

Training histories, learning-rate changes, epochs, stopping status, runtimes, parameter counts, checkpoint sizes, and predictions were archived. Saved checkpoints were replayed to verify reproducibility, and no architecture, optimisation setting, seed, or stopping rule was changed after freezing.

3.5.6 Candidate Screening and Model Freezing

Candidate screening used the predefined primary chronological validation period, with MAE as the principal selection criterion and RMSE as the first tie-breaker when MAE values were similar.

The development registry comprised 19 predefined model–scenario combinations spanning historical baselines, linear and ensemble methods, a multilayer perceptron, recurrent networks, and FT-Transformer. Each candidate used its eligible input representation and fold-local preprocessing, and selection relied only on development-stage chronological validation.

Before locked final-test access, the model registry, feature scenarios, predictor definitions, preprocessing rules, validation assignments, hyperparameters, random seeds, selection criteria, and reporting procedures were frozen. The canonical reporting set comprised the prior two-year moving-average baseline, Elastic Net, FT-Transformer, LSTM, and SimpleRNN. These models were subsequently evaluated under district-grouped, rolling-origin, and locked final-test protocols.

The canonical set represented persistence, regularised linear modelling, tabular attention, gated recurrence, and ungated recurrence, without implying uniform superiority across protocols.

Extra Trees was retained separately for development-stage SHAP feature-attribution and interaction analyses and was not promoted to the canonical predictive comparison. Explainability results were not used for model selection, and excluded candidates were not retuned, replaced, or promoted after locked final-test results became available.

3.6 Validation, Metrics, and Statistical Analysis

The evaluation framework distinguished development-stage chronological screening, future-year forecasting, spatial transfer to districts excluded from model fitting, and one-time assessment on a locked final period. All preprocessing, model fitting, and statistical calculations were performed according to the information restrictions associated with the corresponding validation protocol.

3.6.1 Primary Chronological Validation

The primary chronological validation period comprised crop-year starts 2017, 2018, and 2019. All eligible development observations from crop years preceding 2017 formed the training partition, while the three later crop years formed a contiguous validation block.

The validation block contained 114 observations, corresponding to 38 stable districts observed in each of the three crop years. This later-period evaluation was used to screen the predefined model–scenario candidates and establish the canonical reporting set. The 2017–2019 observations remained part of the development data and were not part of the locked final test.

3.6.2 Rolling-Origin Validation

Rolling-origin validation evaluated future-year forecasting by repeatedly training each model using only crop years preceding the year being predicted. For validation crop year t, the training set was defined as

$$\mathcal{T}_t = \{(d, \tau) : d \in \mathcal{D}, \tau < t, \tau \in \mathcal{E}\}, \tag{21}$$

and the corresponding validation set was

$$\mathcal{V}_t = \{(d, t) : d \in \mathcal{D}\}, \tag{22}$$

where d indexes the stable districts, $\mathcal{D}$ is the set of 38 stable districts, τ denotes an eligible crop year preceding t, and $\mathcal{E}$ is the set of crop years satisfying the historical-window requirements. Thus, no crop year later than or equal to t contributed to model fitting for validation year t.

The completed rolling-origin analysis used nine held-out crop-year starts: 2007, 2008, and 2013-2019. Each validation partition contained all 38 stable districts. Predictor imputation, scaling, feature selection, climatological-reference estimation, anomaly construction, target transformation, and model fitting were repeated independently within every origin.

The nine validation years generated

$$9 \times 38 = 342$$

out-of-fold predictions for each model-level ensemble. This protocol measured performance under temporal distribution shift and represented the study's principal test of future-season forecasting within the development period.

3.6.3 District-Grouped Validation

Spatial transfer was evaluated using five-fold GroupKFold with stable district identity as the grouping variable. Let G_k^{train} and $G_k^{\text{validation}}$ denote the sets of districts assigned to the training and validation partitions in fold k. District separation was enforced as

$$G_k^{\text{train}} \cap G_k^{\text{validation}} = \emptyset. \tag{23}$$

Consequently, every observation from a given district was assigned entirely to either training or validation within a fold. Validation folds contained seven or eight districts, depending on the fold composition.

Across the five folds, all 608 observations in the common development panel received one out-of-fold prediction per model-level ensemble. District identity was used to construct the folds, align outputs, map residuals, and produce district-level summaries, but it was not supplied to the primary models as a predictor. This protocol therefore assessed transfer to districts excluded from model fitting while retaining their eligible historical and meteorological predictor values.

3.6.4 Locked Final Test

Crop-year starts 2020–2022 formed the locked final test, comprising 114 observations from 38 stable districts across three crop years.

Before access, predictors, preprocessing, feature scenarios, architectures, hyperparameters, seeds, validation assignments, statistical procedures, uncertainty rules, and the canonical model set were frozen. Locked observations were excluded from screening, feature selection, climatological estimation, interval calibration, tuning, and model promotion.

Performance was reported both pooled across 2020–2022 and separately by crop year. Year-specific results described temporal variation only and were not used to revise models.

The final test was executed once and then marked as consumed. Subsequent explainability, robustness, reliability, and reporting analyses used saved development-stage outputs, fitted preprocessing pipelines, and frozen checkpoints.

3.6.5 Performance Metrics

All predictions were inverse-transformed and evaluated in kgha^{-1}. Let n denote the number of evaluated district-crop-year observations, y_i the observed yield, $\hat{y}_i$ the predicted yield, and $\bar{y}$ the mean observed yield.

Mean absolute error was the primary metric:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|. \tag{24}$$

MAE represents average prediction error in the original yield unit and was selected because it is directly interpretable and less sensitive to large residuals than squared-error measures.

Root mean-squared error was calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{25}$$

RMSE gives greater weight to large errors and therefore complements MAE by indicating sensitivity to severe prediction failures.

The coefficient of determination was

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{26}$$

R^2 measures the proportion of observed variation represented by the predictions; values closer to one indicate better fit, while negative values indicate performance below the evaluation-set mean.

Mean absolute percentage error was

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \quad (27)$$

MAPE expresses error relative to observed yield and was treated as a supporting metric because percentage errors can become unstable near zero, although all wheat-yield observations were positive.

Mean prediction bias was

$$\text{Bias } = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i). \quad (28)$$

Positive bias indicates systematic overprediction, whereas negative bias indicates underprediction.

Percentage improvement relative to a historical baseline was calculated as

$$\text{Improvement}(\%) = 100\left(\frac{MAE_{\text{baseline}} - MAE_{\text{model}}}{MAE_{\text{baseline}}}\right). \quad (29)$$

Positive values indicate reduced MAE relative to the baseline, zero indicates equivalent performance, and negative values indicate poorer performance. Canonical comparisons used the prior two-year moving-average baseline; comparisons with previous-year persistence were identified explicitly.

Pooled validation metrics were calculated from the complete observation-level prediction set rather than by averaging fold-level metrics. Results were also retained by model, scenario, seed, validation fold, crop year, district, yield range, and outlier status for diagnostic and failure analyses.

3.6.6 Statistical Comparisons

Statistical comparisons used paired prediction errors because competing models predicted the same district-crop-year observations. For models A and B, the paired absolute-error difference was

$$d_i = |y_i - \hat{y}_{i,A}| - |y_i - \hat{y}_{i,B}|, \quad (30)$$

where negative d_i favours model A and positive d_i favours model B. Ninety-five percent confidence intervals for paired performance differences were estimated using 10,000 cluster-level bootstrap resamples. Temporal comparisons were clustered by validation crop year, while spatial comparisons were clustered by stable district, preserving within-cluster dependence.

Rolling-origin comparisons used an exact cluster-level sign-flip test across the nine held-out crop years, evaluating all

$$2^9 = 512$$

possible sign assignments. District-grouped comparisons used 100,000 Monte Carlo sign-flip draws across the 38 district clusters.

Wilcoxon signed-rank tests were used where paired observation-level comparisons were appropriate, with rank-biserial correlation as the associated effect-size measure. Holm correction controlled the family-wise error rate across predefined pairwise comparison families.

Protocol-specific multi-model comparisons used Friedman tests on matched validation blocks, followed by Nemenyi post-hoc comparisons where applicable. Rolling-origin and district-grouped protocols were analysed separately, and no locked final-test observations were used for inferential testing.

Cliff's delta was not added because it largely duplicated the rank-based information from rank-biserial correlation. Cohen's d was not prioritised because paired absolute-error differences could be skewed, while Bayesian model comparison was excluded because it was not part of the frozen statistical plan. No additional statistical procedure was introduced after inspection of the locked final-test results.

3.7 Explainability, Interaction, and Stability Analysis

All explanation analyses were conducted using saved development-stage models, preprocessing objects, predictions, and checkpoints. Models were not retrained for explanation, and observations from the locked final-test period were excluded. The resulting attributions were interpreted as descriptions of fitted-model behaviour rather than as causal estimates of agronomic effects.

3.7.1 Tree-Based

Tree SHAP was applied to the development-selected history-only Extra Trees model using the 114 primary chronological-validation observations. The fitted model and preprocessing pipeline were reused without retraining.

For feature j, the SHAP value was defined as

$$\phi_j(f, x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{j\}) - f_x(S)], \tag{31}$$

where F is the complete predictor set, $M = |F|, S$ is a subset excluding feature j, and $f_x(S)$ is the model output for known features S. The weighting term averages the marginal contribution of feature j across feature orderings.

Global importance was calculated as the mean absolute SHAP value across observations. Observation-level distributions were retained to assess contribution direction and variability, with local explanations retained for individual district-crop-year predictions.

SHAP additivity was verified as

$$\epsilon_i = \left| f(x_i) - \left(\phi_0 + \sum_{j=1}^{M} \phi_{i,j} \right) \right|, \tag{32}$$

where ϕ_0 is the expected model output and ϵ_i the reconstruction discrepancy. Explanations were accepted only within the predefined numerical tolerance. Locked final-test observations were excluded from all SHAP analyses.

3.7.2 SHAP Interaction Values

A separate development-only SHAP interaction analysis was conducted using an Extra Trees model fitted to the selected weather and climate-anomaly predictors. This interaction model was distinct from the historical-only Extra Trees model used for the global SHAP analysis. No locked final-test observations were used.

For two distinct predictors j and k, the SHAP interaction value was defined as

$$\Phi_{j,k}(x) = \sum_{S \subseteq F \setminus \{j,k\}} w(S)[f_x(S \cup \{j,k\}) - f_x(S \cup \{j\}) - f_x(S \cup \{k\}) + f_x(S)], \tag{33}$$

where

$$w(S) = \frac{|S|!\,(M - |S| - 2)!}{2(M-1)!}.$$

The interaction value represents the joint contribution of predictors j and k beyond their separate contributions. Global interaction strength was summarised as the mean absolute interaction value across the explained development observations:

$$\bar{I}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} \left| \Phi_{i,j,k} \right|. \tag{34}$$

The resulting symmetric matrix was displayed as a heatmap, with off-diagonal entries representing pairwise interactions between predictors. Interaction magnitudes were interpreted as dependence within the fitted predictive function and not as evidence of causal or physiological interactions.

3.7.3 Explanation Stability and Cross-Method Agreement

Explanation stability was assessed using development-only attribution outputs without retraining or locked final-test access. For the frozen Elastic Net model, exact linear SHAP values were evaluated using 1,000 district-cluster bootstrap resamples, with confidence intervals, median and extreme ranks, and top-1 and top-3 selection frequencies reported for the eight historical predictors.

For LSTM and SimpleRNN, cross-method agreement compared Integrated Gradients rankings with frozen-model neutralisation rankings for meteorological channels and monthly time steps under district-grouped and rolling-origin validation.

Spearman rank correlation quantified ranking agreement:

$$\rho_s = \mathrm{corr}[R(a), R(b)], \tag{35}$$

where $R(a)$ and $R(b)$ are the corresponding feature or temporal rankings derived from attribution magnitude and neutralisation-induced performance change. Average ranks were assigned for ties, and top-1 agreement and top-3 overlap were also retained.

Cross-model attribution patterns from Elastic Net, FT-Transformer, LSTM, and SimpleRNN were compared descriptively for shared historical predictors. Seed-level SHAP stability and fold-wise SHAP rank stability were not claimed because these analyses were not part of the frozen outputs. All stability and agreement measures were interpreted as fitted-model dependence rather than causal agronomic importance.

3.7.4 Integrated Gradients

Integrated Gradients was used for final development-stage interpretation of the saved LSTM and SimpleRNN models. For input dimension j, attribution was defined as

$$IG_j(x) = \left(x_j - x_j'\right) \int_0^1 \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_j} d\alpha, \tag{36}$$

where x is the observed input, x' the reference input, $F(\cdot)$ the fitted neural model, and α the interpolation parameter.

The reference represented training-centred scaled inputs, equal to zero after fold-local standardisation. Recurrent attributions were aggregated by static historical predictor, meteorological channel, month-channel combination, and monthly position.

Numerical convergence was assessed using 32,64,128,256, and 512 integration points. The audit selected 64-point Gauss-Legendre integration, which was applied to all 84 saved LSTM and SimpleRNN development checkpoints. Attribution completeness was assessed as

$$r(x) = \left| F(x) - F(x') - \sum_{j=1}^{p} IG_j(x) \right|, \tag{37}$$

where p is the number of attributed input dimensions. Lower values indicate closer agreement between the model-output difference and summed attributions.

Integrated Gradients was selected for its suitability to differentiable sequential models, combined weather and historical inputs, and explicit convergence and completeness assessment. DeepSHAP, GradientSHAP, Expected Gradients, LayerCAM, and attention rollout were not benchmarked. Integrated-Gradients outputs were treated as model-specific post hoc explanations, not causal evidence or model-selection criteria.

3.8 Robustness, History–Weather Contribution, and Failure Analysis

Robustness and failure analyses were conducted using saved development-stage predictions, fitted preprocessing objects, and frozen neural checkpoints. No model was retrained solely for these analyses, and locked final-test observations were not accessed. The analyses were intended to determine how strongly the fitted models depended on historical and meteorological information, how their predictions changed when parts of the input were neutralised, and under which observed conditions prediction errors increased.

3.8.1 Matched Historical-Feature Ablation

The matched ablation compared two LSTM configurations: weather-only, using the seven-month meteorological sequence, and weather-plus-history, adding the eight static historical-yield predictors.

Architecture, weather channels, validation assignments, eligible observations, seeds, optimiser, learning-rate schedule, early stopping, target transformation, and fold-local preprocessing were held constant. The historical branch introduced a small number of additional input-related parameters; therefore, the comparison measured the incremental predictive value of historical information within the same recurrent framework rather than a perfectly parameter-matched architecture.

The experiment comprised two feature configurations, three fixed seeds, nine rolling-origin folds, and five district-grouped folds, producing 84 fitted runs and 5,700 observation-level predictions. For observation i, the effect of historical information was measured as

$$\Delta_i^{\text{history}} = \left| y_i - \hat{y}_{i,\,\text{weather + history}} \right| - \left| y_i - \hat{y}_{i,\,\text{weather}} \right|. \tag{38}$$

Negative values indicated lower absolute error after adding historical predictors, whereas positive values favoured the weather-only configuration. Results were summarised using MAE

changes, observation-level win rates, cluster-bootstrap confidence intervals, cluster-level sign-flip tests, and Holm-adjusted probability values.

This ablation measured predictive contribution rather than causality; improvements from historical yield did not identify specific soil, irrigation, management, or technological drivers.

3.8.2 Frozen-Model Feature Neutralisation

Frozen-model feature neutralisation assessed how fitted Elastic Net, FT-Transformer, LSTM, and SimpleRNN models responded when selected inputs were replaced by neutral values. The prior two-year moving-average model was retained as a historical reference where appropriate. Three perturbations were examined: neutralisation of the complete historical-yield feature block, individual monthly meteorological channels, and individual monthly time steps. For standardised predictors, zero represented the training-partition centre. Preprocessing objects, model parameters, fold assignments, random seeds, and tensor dimensions remained unchanged; models were replayed without retraining.

For model m, the prediction shift produced by perturbation q was defined as

$$\delta_{i,m,q} = \hat{y}_{i,m}^{(q)} - \hat{y}_{i,m}^{\text{reference}}, \tag{39}$$

where $\hat{y}_{i,m}^{(q)}$ is the perturbed prediction and $\hat{y}_{i,m}^{\text{reference}}$ is the original prediction. The absolute shift measured sensitivity magnitude, while the signed shift indicated the direction of prediction change.

Each fitted pipeline or neural checkpoint was first replayed on unmodified inputs, and perturbations were accepted only when archived canonical predictions were reproduced within the predefined numerical tolerance. The analysis included 44 active perturbations under district-grouped and rolling-origin validation.

Perturbed and reference predictions were compared using changes in MAE, RMSE, R^2, MAPE, and mean bias, together with signed and absolute prediction shifts, cluster-bootstrap confidence intervals, sign-flip tests, and Holm correction. Locked final-test data were excluded.

Feature neutralisation measured dependence within fitted models; it was not equivalent to retraining without the affected feature group and did not establish causality.

3.8.3 Sequence-Length Sensitivity

Sequence-length sensitivity was evaluated by retaining the most recent one to seven months of the October–April sequence while replacing earlier positions with training-centred neutral values and preserving the original seven-position tensor.

The one-month configuration retained April only; subsequent configurations progressively added March, February, January, December, November, and October, with the seven-month configuration retaining the full sequence.

Architecture, fitted weights, static historical branch, scalers, validation folds, and random seeds were unchanged. All 84 authoritative recurrent checkpoints were evaluated, and full-sequence replay was required to reproduce archived predictions within numerical tolerance before shortened-sequence results were accepted.

Sensitivity was assessed using pooled MAE, changes relative to the full sequence, cluster-bootstrap confidence intervals, cluster-level sign-flip tests, and Holm-adjusted probability values.

This analysis measured sensitivity of already fitted full-season models to retained monthly information. It was not equivalent to training operational early-season forecasting models, because the models were originally fitted on complete October–April sequences and neutralised future positions did not reproduce genuine early-season retraining conditions.

3.8.4 Error and Failure-Mode Analysis

For observation i, the signed residual was defined as

$$r_i = \hat{y}_i - y_i, \tag{40}$$

where $\hat{y}_i$ is predicted yield and y_i observed yield. Positive and negative residuals indicate overprediction and underprediction, respectively. Absolute error was

$$e_i = |\hat{y}_i - y_i|. \tag{41}$$

Failure analysis examined residual and absolute-error distributions, district- and crop-year-level MAE, observed-yield ranges, high-yield outliers, and temporal error patterns.

Potential outliers were identified using the Tukey 1.5-interquartile-range rule from development-stage training data. The frozen limits were approximately -495.45 and 4,576.92 kgha^{-1}; because yields were non-negative, only the upper limit was operationally relevant. Five high-yield observations were identified under each development protocol and retained in the principal analysis, with performance also reported for the complete, non-outlier, and high-yield subsets.

Temporal drift was assessed using chronological crop-year residual and MAE trajectories without modifying the fitted models.

Prediction agreement was additionally assessed using Bland-Altman analysis. For observation i, the model-observation difference and pairwise mean were

$$d_i = \hat{y}_i - y_i, a_i = \frac{\hat{y}_i + y_i}{2}. \quad (42)$$

Mean prediction bias was

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad (43)$$

and the approximate 95% limits of agreement were

$$\bar{d} \pm 1.96 s_d, \quad (44)$$

where s_d is the sample standard deviation of the differences. Bland-Altman plots assessed bias, dispersion, and disagreement across the yield range.

Late-sowing analysis was unavailable because consistent sowing-date data were absent. Drought-, heatwave-, and extreme-rainfall-related failure claims were made only where predefined subgroup results were available; feature attribution alone was not treated as evidence of failure under a weather regime.

3.8.5 Learning-Curve and Overfitting Diagnostics

Learning-curve diagnostics used archived neural-model optimisation histories without retraining. Records included epoch-level training loss, validation MAE, learning-rate changes, selected epoch, early-stopping status, completed epochs, checkpoint restoration, fold identity, validation year, and random seed.

Training and validation trajectories were examined jointly. Declining training loss with worsening validation MAE was treated as evidence consistent with overfitting, while similar trajectories followed by stabilisation indicated convergence. Irregular fold- or seed-specific patterns were interpreted as optimisation instability.

Learning-rate changes were compared with validation trajectories to assess scheduler effects, and selected checkpoint epochs were verified against the minimum recorded validation MAE rather than the final training epoch. Seed- and fold-level histories were retained separately to distinguish architectural behaviour from variation due to initialisation, validation composition, or prediction year.

These diagnostics were used to examine protocol-specific differences between SimpleRNN and LSTM, including possible optimisation variability or training–validation divergence on the seven-month sequence. SimpleRNN was not assumed to be universally superior because performance differed across district-grouped, rolling-origin, and locked final-test evaluations. Learning curves were therefore interpreted alongside parameter count, validation protocol,

seed variation, and prediction errors. The analyses were diagnostic only and did not alter frozen architectures, stopping rules, the canonical model set, or consumed locked final-test results.

3.9 Prediction Uncertainty and Calibration

Prediction reliability was evaluated for the five canonical models: the prior two-year moving-average baseline, Elastic Net, FT-Transformer, LSTM, and SimpleRNN. The analysis used saved development-stage out-of-fold predictions and did not access or recalibrate against locked final-test observations. A common residual-based procedure was applied to all models so that interval reliability and sharpness could be compared under identical spatial and temporal validation structures.

3.9.1 Cluster-Excluded Prediction Intervals

Prediction intervals were constructed using cluster-excluded absolute out-of-fold residuals. For observation i , the absolute calibration residual was

$$a_i = |y_i - \hat{y}_i|, \tag{45}$$

where y_i is the observed yield and $\hat{y}_i$ the corresponding out-of-fold prediction. For nominal coverage $1 - \alpha$, the symmetric prediction interval was

$$I_{i,1-\alpha} = \left[\hat{y}_i - q_{m,p,c,1-\alpha}, \hat{y}_i + q_{m,p,c,1-\alpha}\right], \tag{46}$$

where $q_{m,p,c,1-\alpha}$ is the empirical $1 - \alpha$ quantile of cluster-excluded absolute residuals for model m, validation protocol p, and excluded cluster c. The $1 - \alpha$ quantile was used because calibration relied on absolute residuals; using $q_{1-\alpha/2}$ would produce unnecessarily wider intervals.

For district-grouped validation, all records from the target district were excluded, leaving

$$608 - 16 = 592$$

calibration residuals. For rolling-origin validation, all observations from the target crop year were excluded, leaving

$$342 - 38 = 304$$

This prevented the target district or crop year from contributing to its own interval calibration. Separate residual quantiles were estimated for nominal coverage levels of 50%, 80%, 90%, and 95%.

These intervals were development-stage prediction-reliability summaries, not causal uncertainty bounds, posterior credible intervals, or guarantees of coverage for unrelated regions.

3.9.2 Calibration and Reliability Metrics

Let L_i and U_i denote the lower and upper prediction-interval limits for observation i. Empirical coverage at nominal level $1-\alpha$ was

$$\hat{C}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \in [L_i, U_i]), \tag{47}$$

where $1(\cdot)$ is the indicator function. Signed coverage deviation was

$$D_{1-\alpha} = \hat{C}_{1-\alpha} - (1-\alpha), \tag{48}$$

and absolute coverage deviation was

$$|D_{1-\alpha}| = \left|\hat{C}_{1-\alpha} - (1-\alpha)\right|. \tag{49}$$

Positive and negative signed deviations indicated overcoverage and undercoverage, respectively.

Mean interval width was

$$MIW = \frac{1}{n}\sum_{i=1}^{n} (U_i - L_i). \tag{50}$$

Mean and median interval widths were retained to assess sharpness alongside coverage. The interval score for nominal coverage $1-\alpha$ was

$$IS_{\alpha,i} = (U_i - L_i) + \frac{2}{\alpha}(L_i - y_i)1(y_i < L_i) + \frac{2}{\alpha}(y_i - U_i)1(y_i > U_i), \tag{51}$$

with mean interval score

$$\overline{IS}_{\alpha} = \frac{1}{n}\sum_{i=1}^{n} IS_{\alpha,i}. \tag{52}$$

Lower scores indicated a better balance between coverage and sharpness. Below- and above-interval rates were

$$R_{\text{below}} = \frac{1}{n}\sum_{i=1}^{n} 1(y_i < L_i), \tag{53}$$

and

$$R_{\text{above}} = \frac{1}{n}\sum_{i=1}^{n} 1(y_i > U_i). \tag{54}$$

These measures identified systematic interval displacement relative to observations. Reliability was assessed using empirical coverage, signed and absolute coverage deviation, mean and median interval width, mean interval half-width, interval score, and below- and above-interval rates. Nominal and empirical coverage were compared graphically against the 45° ideal-calibration line.

Conditional reliability was also examined by model, validation protocol, district, crop year, observed-yield range, and outlier status. Coverage and interval width were interpreted jointly because high coverage with excessively wide intervals indicates poor sharpness.

The terms expected calibration error and uncertainty calibration error were not used because separate ECE or UCE formulations were not calculated; reliability was quantified directly using the measures defined above.

3.9.3 Scope of the Uncertainty Analysis

A common cluster-excluded residual-calibration procedure was used for all five canonical models to preserve protocol-specific clustering and enable direct comparison of empirical coverage, interval width, and interval score without architecture-specific retraining.

Monte Carlo dropout, deep ensembles, quantile regression, and alternative conformal methods were not benchmarked. Monte Carlo dropout was not uniformly applicable across model classes, deep ensembles would have required additional training beyond the frozen seed design, and quantile regression or alternative conformal methods would have required new objectives or calibration choices after model freezing.

The resulting intervals are therefore described as cluster-excluded empirical residual intervals. They do not constitute a comprehensive uncertainty-quantification benchmark or demonstrate superiority over alternative methods. Comparative evaluation of these approaches remains for future externally validated work.

3.10 Sentinel-2 Seasonal Association Analysis

The Sentinel-2 component provided a complementary assessment of district-level vegetation and moisture conditions during recent wheat seasons. It was conducted separately from the principal prediction benchmark because consistent Sentinel-2 Level-2A coverage was available only for the six crop years from 2017–2018 to 2022–2023. Satellite-derived variables were linked with official district-level agricultural statistics for descriptive association analysis.

3.10.1 Seasonal Windows

Three fixed crop-stage windows were defined for each Rabi wheat season:

- Early season: 1 November–31 December;
- Peak growth: 1 January–the final day of February;
- Late season: 1 March–30 April.

These calendar-based windows broadly represented crop establishment, canopy development, and late-season decline or senescence and were applied consistently across all districts and crop years because district–crop-year phenological dates were unavailable.

Early-to-peak differences represented vegetation establishment and canopy development, while peak-to-late differences represented subsequent decline or senescence. Missing stage-level satellite values were retained as missing and not temporally interpolated.

3.10.2 Image and Cloud Processing

Harmonised Sentinel-2 Level-2A surface-reflectance imagery was obtained from the COPERNICUS/S2_SR_HARMONIZED collection to maintain numerical consistency across processing baselines. Reflectance values were multiplied by 0.0001 before spectral-index calculation.

Sentinel-2 scenes were linked with GOOGLE/CLOUD_SCORE_PLUS/V1/S2_HARMONIZED observations using common image identifiers. Pixels were retained only when the cs_cdf optical-quality score was ≥0.60.

The Scene Classification Layer was also used to exclude classes 0, 1, 3, 8, 9, 10, and 11, representing no data, defective pixels, cloud shadows, medium- and high-probability clouds, cirrus, and snow or ice.

After filtering, valid observations within each seasonal window were combined using median compositing to reduce residual contamination and anomalous reflectance effects.

3.10.3 Cropland Mask and Spectral Indices

ESA WorldCover 2021 v200 class 40 was used as the fixed general-cropland mask for the final district-level statistics. Dynamic World V1 crop probability was examined as supporting land-cover information but was not used to redefine the frozen publication mask.

District–crop-year records were retained only when the mapped-cropland fraction satisfied the predefined minimum threshold of 0.10. The mask represented mapped general cropland and was not interpreted as a verified wheat-specific classification.

Four spectral indices were calculated for the early, peak, and late seasonal composites. The Normalised Difference Vegetation Index was calculated as

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}}, \tag{55}$$

where ρ_{NIR} and ρ_{Red} denote near-infrared and red surface reflectance, respectively. NDVI represented broad differences in vegetation greenness and canopy activity.

The Enhanced Vegetation Index was calculated as

$$\mathrm{EVI} = 2.5\frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + 6\rho_{\mathrm{Red}} - 7.5\rho_{\mathrm{Blue}} + 1}, \tag{56}$$

where ρ_{Blue} denotes blue-band reflectance. EVI was included to provide a vegetation measure with additional atmospheric and soil-background adjustment.

The Soil-Adjusted Vegetation Index was calculated as

$$\mathrm{SAVI} = (1 + L)\frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}} + L}, L = 0.5, \quad (57)$$

where L is the soil-adjustment constant. SAVI was used to reduce the influence of exposed soil where vegetation cover was incomplete.

The Normalised Difference Moisture Index was calculated as

$$\mathrm{NDMI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{SWIR1}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{SWIR1}}}, \quad (58)$$

where ρ_{SWIR1} denotes short-wave-infrared reflectance. NDMI represented relative differences in canopy and surface moisture conditions.

For each index, early-to-peak and peak-to-late changes were calculated as

$$\Delta I_{\text{early} \rightarrow \text{peak}} = I_{\text{peak}} - I_{\text{early}},$$

and

$$\Delta I_{\text{peak} \rightarrow \text{late}} = I_{\text{late}} - I_{\text{peak}},$$

where $\boldsymbol{I}$ denotes NDVI, EVI, SAVI, or NDMI.

3.10.4 District-Level Aggregation

District summaries were generated from validated seasonal image stacks at the frozen publication-statistics scale of 500 m. For each spectral index and seasonal stage, retained statistics included valid-pixel count, mean, median, standard deviation, minimum, maximum, 25th and 75th percentiles, interquartile range, and coefficient of variation.

The final satellite panel contained 228 district–crop-year observations, representing 38 stable districts across six crop years. Each record was aligned with stable district, crop-year start, official cultivated wheat area, wheat production, and reconstructed wheat yield.

Thirty publication maps were generated: early-season, peak-growth, and late-season NDVI, together with early-to-peak and peak-to-late NDVI change, for each of the six crop years. Maps were derived from the same validated district-level seasonal assets used for the tabular satellite statistics.

3.10.5 Satellite-Yield Association

Pearson and Spearman correlations were calculated between official district-level wheat yield and stage-specific or seasonal-change satellite variables. Pearson correlation measured linear association, while Spearman correlation measured monotonic rank association.

Tested variables included early-, peak-, and late-stage NDVI, EVI, SAVI, and NDMI, together with early-to-peak and peak-to-late differences. Exact valid sample sizes were retained because

cloud filtering, cropland masking, or missing seasonal composites could reduce observations for individual comparisons.

Associations were evaluated across the complete six-year district panel and interpreted by direction, magnitude, valid sample size, and consistency between Pearson and Spearman estimates.

3.10.6 Interpretation Boundary

The Sentinel-2 analysis was descriptive and complementary, representing district-level associations between official wheat yield and spectral conditions over mapped general cropland.

These associations were not interpreted causally and could reflect temporal trends, district productivity, irrigation, management, soil conditions, mixed cropping, or other unmeasured factors. The cropland mask was not wheat-specific; therefore, the analysis did not estimate wheat classification accuracy, wheat-specific area, or field-level yield.

Satellite variables were not part of the principal long-period prediction benchmark, which relied on agricultural history and ERA5-Land meteorology available consistently across the full study period. Sentinel-2 instead provided recent-period contextual evidence on relationships between district-level wheat productivity and seasonal general-cropland spectral conditions.

3.11 Reproducibility, Data Governance, and Ethics

3.11.1 Reproducibility and Computational Reporting

The analysis used Windows and Python 3.11.0 with pandas 2.2.3, NumPy 1.26.4, GeoPandas 1.1.4, Shapely 2.0.7, scikit-learn 1.8.0, Rasterio 1.4.4, PyTorch 2.11.0+cu126, XGBoost 2.1.4, LightGBM 4.6.0, and CatBoost 1.2.10.

CUDA was available with one device and PyTorch compiled against CUDA 12.6, but CPU was used as the controlled primary training device. Deterministic PyTorch and cuDNN operations were enabled, with cuDNN benchmarking disabled.

Reproducibility controls included fixed seeds (20260721, 20260722, and 20260723), deterministic district and crop-year ordering, frozen feature and model registries, archived fold assignments, preprocessing objects, fitted pipelines, neural checkpoints, training histories, predictions, manifests, and cryptographic checksums. The locked 2020–2022 final test was executed once and marked as consumed; subsequent analyses used saved development-stage artifacts without altering the frozen workflow.

Computational reporting included training and prediction runtime, trainable parameter count, fitted-model or pipeline size, and neural-checkpoint size. Because runtimes were recorded across different execution stages, they were treated as descriptive rather than hardware-normalised benchmarks and interpreted alongside predictive accuracy, model complexity, and storage requirements.

3.11.2 Data Governance and Ethics

The study used only secondary, non-personal agricultural, meteorological, administrative-boundary, land-cover, and satellite data aggregated at district level. No human participants, animal experiments, personally identifiable information, or confidential farm-level records were involved. Model predictions, explanations, and satellite associations were interpreted only at the aggregated research level and not as causal assessments of individual farmers, management practices, or institutions.

4. Results

4.1 Data Harmonisation and Analytical Coverage

4.1.1 Agricultural and Administrative Data

The agricultural extract contained 1,217 Rabi wheat records covering 52 compatible district names and 25 available crop years, while the contemporary boundary layer contained 55 district polygons. Name standardisation, geometry repair, historical crosswalking, and polygon dissolution produced 38 stable analytical districts.

Aggregation of cultivated area and production yielded 950 unique district–crop-year observations across the 38 districts and 25 crop years. Yield was recalculated from aggregated production and area. Final harmonisation identified no duplicate keys, unmatched units, invalid geometries, or unresolved target inconsistencies.

4.1.2 Development, Validation, and Locked-Test Panels

Historical-feature eligibility produced a common development panel of 608 district–crop-year observations from 38 stable districts across 16 crop years. The primary chronological-validation period, crop-year starts 2017–2019, contained 114 observations.

District-grouped validation generated 608 out-of-fold predictions per model across five GroupKFold partitions, with each observation evaluated once and no district shared between training and validation within a fold. Rolling-origin validation generated 342 out-of-fold predictions per model across nine held-out crop years, with 38 districts per year.

The locked 2020–2022 final-test panel contained 114 observations, with all 38 districts represented annually. It was executed once after the model registry, preprocessing, validation assignments, statistical plan, and uncertainty procedures were frozen.

4.1.3 Missing-Year and Eligibility Effects

The complete 2009–2010 agricultural record was unavailable and was not reconstructed or imputed. Crop-year starts 1997–1999 were retained only for historical-feature initialisation.

Because historical predictors required exact preceding crop-year information, crop-year starts 2010–2012 were excluded from matched modelling because their historical windows intersected the missing 2009–2010 record. Eligible development observations therefore came from 2000–2008 and 2013–2019, yielding 608 observations while preserving temporal integrity and avoiding synthetic historical yields.

4.1.4 Weather, Sequence, and Satellite Coverage

The seasonal ERA5-Land dataset contained 950 district–crop-year records, matching the harmonised agricultural panel. The monthly dataset contained 6,650 district–month records, representing seven October–April observations for each district–crop-year combination.

Each recurrent input contained all seven months in the prescribed order, with 13 meteorological channels per month and eight static historical-yield predictors. Final checks found no disagreement between tabular and sequential targets and no non-finite values in accepted inputs.

The complementary Sentinel-2 panel contained 228 district–crop-year records from 38 stable districts across six crop years (2017–2018 to 2022–2023). It included early-, peak-, and late-stage NDVI, EVI, SAVI, and NDMI summaries and early-to-peak and peak-to-late changes. Variable-specific valid sample sizes were retained where quality filtering reduced observations. Final analytical coverage is summarised in Table 7.

Table 7. Final analytical datasets and evaluation coverage

Data component	Final size	Spatial coverage	Temporal coverage	Audit or analytical outcome
Original agricultural extract	1,217 records	52 source-compatible district names	25 available crop years	Source audit completed

Contemporary boundary layer	55 polygons	Madhya Pradesh	Contemporary geography	Geometry and identifier audit completed
Stable agricultural panel	950 observations	38 stable districts	25 available crop years	No duplicate or unmatched final units
Seasonal ERA5-Land panel	950 observations	38 districts	25 available crop years	Complete district–crop-year alignment
Monthly ERA5-Land panel	6,650 records	38 districts	Seven months per district–crop-year	Complete October–April sequences
Common development panel	608 observations	38 districts	16 eligible crop years	Complete historical-feature windows
Primary chronological validation	114 observations	38 districts	2017–2019	Candidate-screening panel
District-grouped validation	608 OOF predictions per model	Five district-grouped folds	16 development crop years	Each development observation evaluated once
Rolling-origin validation	342 OOF predictions per model	38 districts per held-out year	Nine validation crop years	Chronological ordering preserved
Locked final test	114 observations	38 districts	2020–2022	One-time evaluation completed
Sentinel-2 panel	228 observations	38 districts	Six crop years, 2017–2022 starts	District–crop-year panel available for complementary association analysis

4.2 Candidate Screening and Development-Stage Performance

4.2.1 Primary Chronological Screening

Primary chronological screening evaluated 19 model–scenario combinations on 114 observations from crop-year starts 2017–2019, spanning historical baselines, linear and ensemble models, a multilayer perceptron, recurrent networks, and FT-Transformer. Tabular finalists used the development-selected history-only representation, while recurrent finalists combined the seven-month weather sequence with eight static historical-yield predictors. Locked final-test observations were excluded.

The lowest validation MAEs were obtained by ordinary least squares (396.613 kg ha^{-1}), Ridge (396.613 kg ha^{-1}), Elastic Net (396.640 kg ha^{-1}), FT-Transformer (402.176 kg ha^{-1}), and the multilayer perceptron (402.887 kg ha^{-1}). The prior two-year moving average was the best historical baseline at 407.741 kg ha^{-1}. Recurrent-model MAEs were 504.620, 536.463, and 561.284 kg ha^{-1} for LSTM, GRU, and SimpleRNN, respectively. Table 8 presents the complete chronological-validation MAE and RMSE results and the subsequent analytical roles of all 19 candidates.

Table 8. Primary chronological-validation performance and analytical roles of the 19 frozen development finalists

Model family	Model	Development-selected input scenario	Validation MAE (kg ha^{-1})	Validation RMSE (kg ha^{-1})	Subsequent analytical role
Historical baseline	Overall training mean	Historical/training statistics	1,317.036	1,496.832	Development-stage comparator
Historical baseline	District historical mean	Historical/training statistics	1,289.292	1,403.872	Development-stage comparator
Historical baseline	Previous-year persistence	Historical only	438.114	637.706	Development-stage comparator
Historical baseline	Prior two-year moving average	Historical only	407.741	572.218	Canonical baseline

Historical baseline	Prior three-year moving average	Historical only	423.340	564.983	Development-stage comparator
Linear	Ordinary least squares	Historical only	396.613	549.098	Development-stage comparator
Linear	Ridge	Historical only	396.613	549.098	Development-stage comparator
Linear	Elastic Net	Historical only	396.640	549.142	Canonical classical model
Bagging	Random Forest	Historical only	441.893	598.401	Development-stage comparator
Bagging	Extra Trees	Historical only	426.301	578.127	Tree-based explainability model
Boosting	HistGradientBoosting	Historical only	444.749	589.231	Development-stage comparator
Boosting	XGBoost	Historical only	461.300	612.664	Development-stage comparator
Boosting	LightGBM	Historical only	430.829	579.881	Development-stage comparator
Boosting	CatBoost	Historical only	442.742	602.237	Development-stage comparator
Feed-forward neural	Multilayer perceptron	Historical only	402.887	554.523	Development-stage comparator

Recurrent neural	GRU	Monthly weather + static history	536.463	659.815	Development-stage comparator
Recurrent neural	LSTM	Monthly weather + static history	504.620	635.041	Canonical recurrent model
Recurrent neural	SimpleRNN	Monthly weather + static history	561.284	682.065	Canonical recurrent model
Attention-based neural	FT-Transformer	Historical only	402.176	551.103	Canonical tabular deep model

Screening results were interpreted alongside the predefined analytical roles of each model family rather than MAE ranking alone. The canonical set comprised the prior two-year moving-average baseline, Elastic Net, FT-Transformer, LSTM, and SimpleRNN, while Extra Trees was retained separately for explainability. Other candidates remained development-stage evidence and were not reconsidered after locked final-test inspection. Table 9 summarises the five canonical models and their input representations.

Table 9. Canonical reporting models and input representations

Model	**Model class**	**Input representation**	**Principal reporting role**
Prior two-year moving average	Historical baseline	Prior two-year yield mean	Persistence benchmark
Elastic Net	Regularised linear model	Eight historical-yield predictors	Canonical classical model
FT-Transformer	Tabular attention model	Eight historical-yield predictors	Canonical tabular deep model
LSTM	Gated recurrent neural network	Seven-month weather sequence + eight static historical predictors	Canonical gated recurrent model and ablation backbone

SimpleRNN	Recurrent neural network	Seven-month weather sequence + eight static historical predictors	Canonical ungated recurrent model

4.2.2 District-Grouped Validation

District-grouped validation assessed geographical transfer to districts excluded from model fitting while retaining eligible historical context. SimpleRNN achieved the lowest pooled MAE (304.208 kg ha^{-1}), followed by LSTM (307.407), Elastic Net (311.687), the prior two-year moving-average baseline (312.902), and FT-Transformer (320.752 kg ha^{-1}).

SimpleRNN improved upon the baseline by 8.694 kg ha^{-1} (2.78%) and differed from LSTM by only 3.199 kg ha^{-1} (approximately 1.04%). The total MAE range across the five models was 16.544 kg ha^{-1}.

Among deep models, SimpleRNN also achieved the lowest RMSE (411.512 kg ha^{-1}), highest (0.831), and lowest MAPE (14.642%). LSTM recorded RMSE 415.320 kg ha^{-1}, MAPE 15.010%, and bias −5.052 kg ha^{-1}, while FT-Transformer recorded RMSE 438.043 kg ha^{-1}, MAPE 15.709%, and bias −11.284 kg ha^{-1}. Table 10 compares canonical-model MAEs under district-grouped and rolling-origin validation.

Table 10. Canonical model MAE under district-grouped and rolling-origin validation

Model	**District-grouped MAE (kgha^{-1})**	**Rolling-origin MAE (kg ha^{-1})**	**Rolling-origin minus districtgrouped MAE (kgha^{-1})**
Prior two-year moving average	312.902	370.662	57.760
Elastic Net	311.687	389.087	77.400
FT-Transformer	320.752	367.610	46.858
LSTM	307.407	430.873	123.466
SimpleRNN	304.208	459.834	155.626

Figure 2 presents the district-grouped MAE of the five canonical models. SimpleRNN and LSTM recorded the two lowest errors, while FT-Transformer recorded the highest. The narrow overall range indicates comparatively similar pooled performance under geographical transfer.

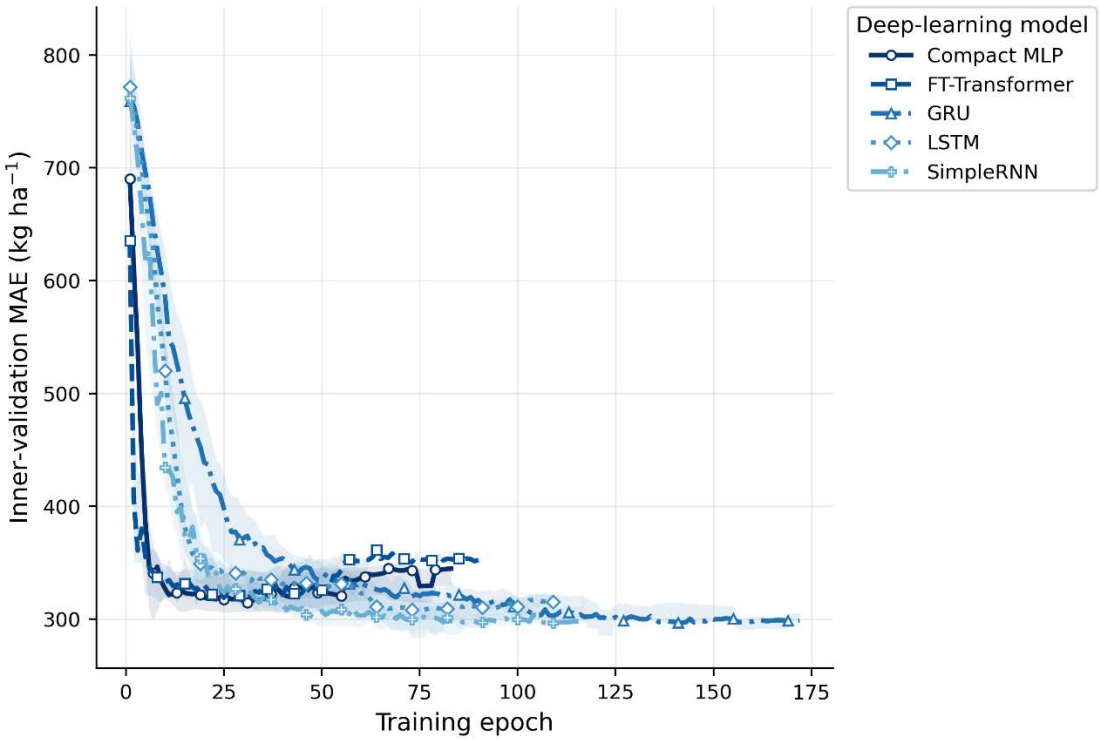


Figure 2. District-grouped validation MAE of the canonical models.

4.2.3 Rolling-Origin Validation

Rolling-origin validation produced a different ranking. FT-Transformer achieved the lowest pooled MAE (367.610 kg ha⁻¹), followed by the prior two-year moving-average baseline (370.662) and Elastic Net (389.087 kg ha⁻¹). FT-Transformer improved upon the baseline by 3.052 kg ha⁻¹ (0.82%).

LSTM and SimpleRNN recorded MAEs of 430.873 and 459.834 kg ha⁻¹, respectively, giving LSTM a 28.961 kg ha⁻¹ advantage and reversing their district-grouped ordering. All models performed worse under rolling-origin validation, with MAE increases ranging from 46.858 kg ha⁻¹ for FT-Transformer to 155.626 kg ha⁻¹ for SimpleRNN.

Supporting recurrent-model metrics showed similar temporal degradation. LSTM recorded RMSE 567.470 kg ha⁻¹, MAPE 15.574%, and bias −224.799 kg ha⁻¹; SimpleRNN recorded RMSE 588.858 kg ha⁻¹, MAPE 16.494%, and bias −253.354 kg ha⁻¹. Table 11 reports the complete supporting metrics for the principal deep models.

Table 11. Supporting performance metrics for the principal deep models

Protocol	Model	RMSE (kg ha^{-1})	R^2	MAPE (%)	Mean bias (kg ha^{-1})

District grouped	FT-Transformer	438.043	0.808	15.709	-11.284
District grouped	LSTM	415.320	0.828	15.010	-5.052
District grouped	SimpleRNN	411.512	0.831	14.642	-8.925
Rolling origin	LSTM	567.470	0.648	15.574	-224.799
Rolling origin	SimpleRNN	588.858	0.621	16.494	-253.354

Figure 3 presents the rolling-origin MAE of the canonical models. FT-Transformer and the persistence baseline occupied the two lowest-error positions, whereas the recurrent models showed larger future-year errors. Model separation was greater than under district-grouped validation.

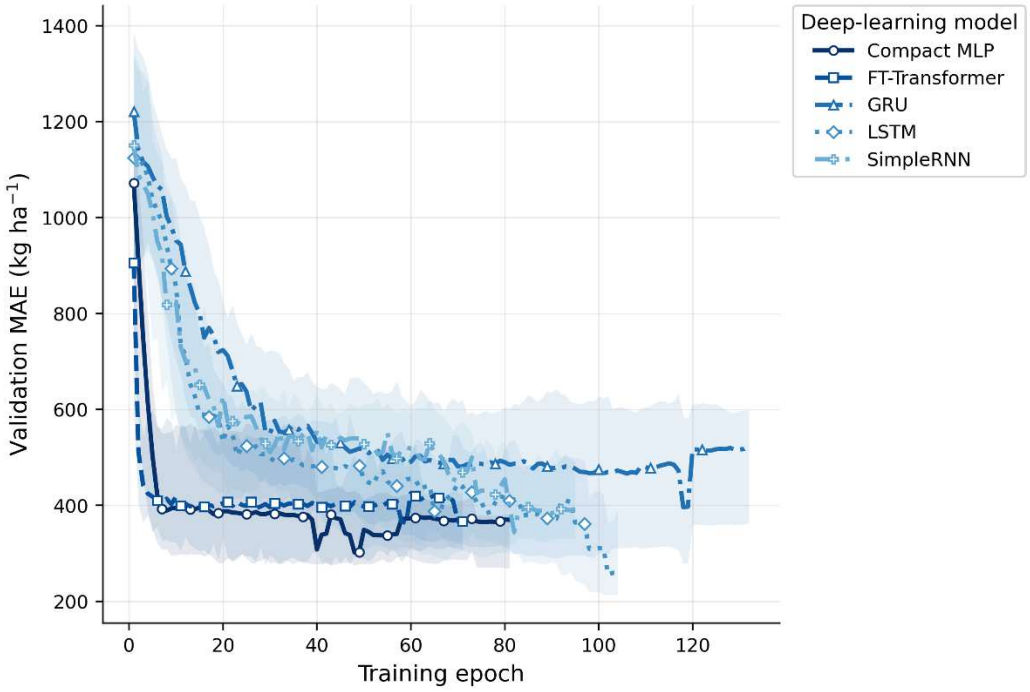


Figure 3. Rolling-origin validation MAE of the canonical models.

The cross-protocol differences are further summarised in Figure 4 using a Taylor diagram. The figure jointly represents predictive correlation, variability, and error structure and provides a visual comparison of model behaviour under district-grouped and rolling-origin validation. It should be interpreted together with the numerical performance metrics rather than as an additional statistical test.

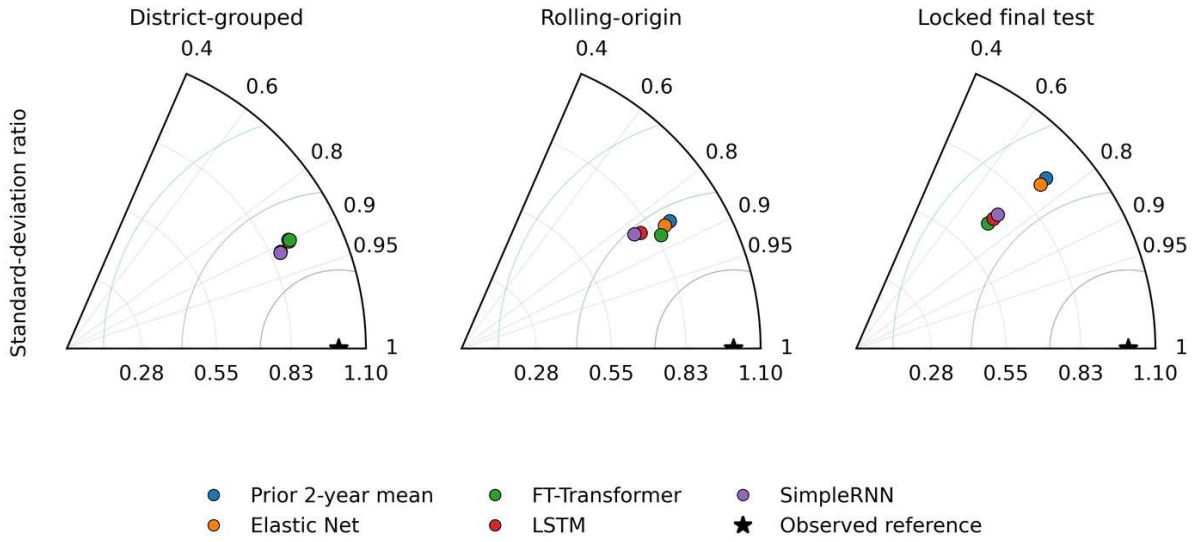


Figure 4. Taylor-diagram comparison of the five canonical models under district-grouped and rolling-origin validation.

Figure 5 presents the corresponding radar-chart comparison of the canonical models across the reported performance metrics. The radar profiles provide a multidimensional visual summary of relative model behaviour across the two validation protocols and should be interpreted alongside the original MAE, RMSE, R^2, MAPE, and mean-bias values.

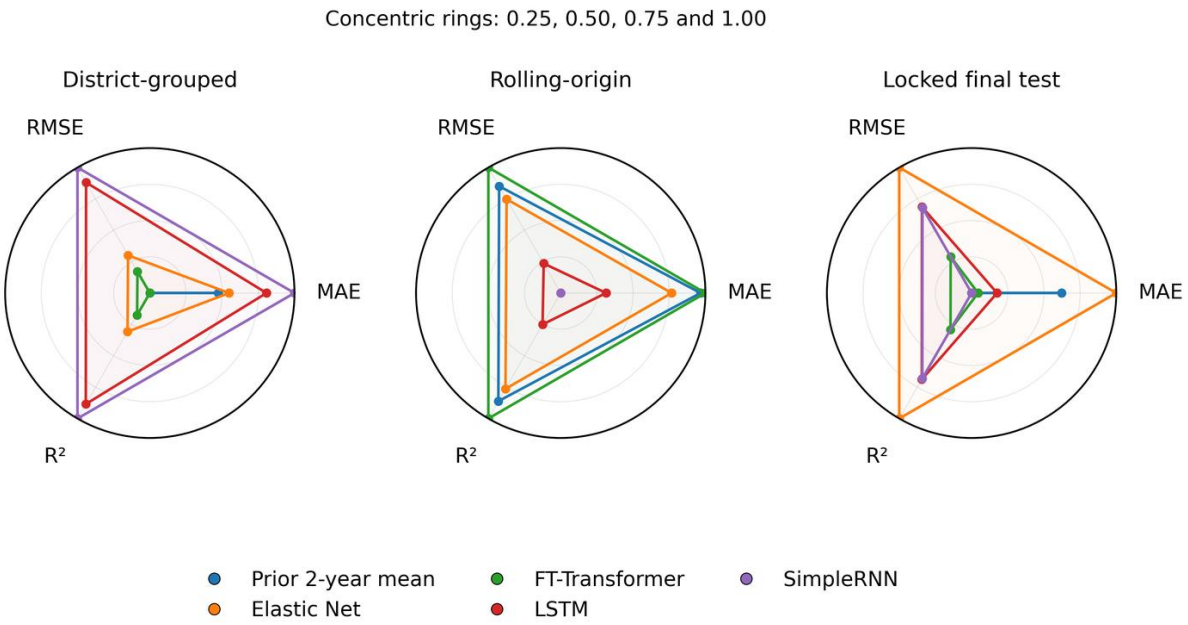


Figure 5. Radar-chart comparison of canonical-model performance under district-grouped and rolling-origin validation.

4.2.4 Statistical Comparison of Canonical Models

Under district-grouped validation, the Friedman test found no significant overall difference among the five canonical models. Elastic Net and the prior two-year moving-average baseline differed by only 1.216 kgha^{-1}, with a rank-biserial effect of 0.045, while the SimpleRNN-LSTM difference was also not statistically resolved. No pairwise comparison remained significant after Holm correction.

Under rolling-origin validation, the Friedman test indicated a significant overall model difference, and Nemenyi analysis identified a significant rank difference between FT-Transformer and SimpleRNN. However, no paired Wilcoxon comparison remained significant after Holm correction. Elastic Net versus FT-Transformer showed a large rank-biserial effect of 0.956, but the Holm-adjusted p-value was 0.0781.

Overall, statistical results supported protocol-dependent numerical rankings rather than consistent superiority of a single model across spatial and temporal validation.

4.2.5 Training and Reproducibility Audit

All neural evaluations used fixed seeds 20260721, 20260722, and 20260723, with model-level predictions obtained by averaging eligible seed outputs rather than selecting the best seed. Final development-stage training durations were 27 epochs for FT-Transformer, 49 for LSTM, and 51 for SimpleRNN.

All 42 expected best-epoch records were available for LSTM and SimpleRNN. FT-Transformer had 41 of 42 records, and the predefined sensitivity rule supported 27 epochs without reconstructing the missing record. Training histories confirmed early stopping, learning-rate scheduling, gradient clipping, and best-checkpoint restoration.

All 126 authoritative deep-model checkpoints were successfully replayed, reproducing archived validation predictions with a maximum difference of 0.0. The canonical model set, configurations, preprocessing, and seed policy were frozen before final-test evaluation, and no candidate was subsequently retuned or promoted.

4.3 Locked Final-Test Performance

4.3.1 Complete Pooled Performance

The locked final test contained 114 district-crop-year observations from crop-year starts 2020, 2021, and 2022, with all 38 stable districts represented in each year. Elastic Net achieved the lowest pooled MAE of $368.498\ \text{kgha}^{-1}$, followed by the prior two-year moving-average baseline at $382.996\ \text{kgha}^{-1}$. LSTM recorded the lowest MAE among the deep models at $400.087\ \text{kgha}^{-1}$, followed by FT-Transformer at $405.125\ \text{kgha}^{-1}$ and SimpleRNN at $406.841\ \text{kgha}^{-1}$.

Elastic Net also recorded the lowest RMSE at $599.288\ \text{kgha}^{-1}$ and the highest R^2 at 0.532. The three deep models showed comparatively similar pooled RMSE and R^2 values, despite their differences in MAE. MAPE ranged from 12.106% for Elastic Net to 13.168% for SimpleRNN. Mean bias was positive for Elastic Net, LSTM, and SimpleRNN and negative for the prior two-year moving-average baseline and FT-Transformer. The complete pooled final-test performance is presented in Table 12.

Table 12. Locked final-test performance of the five canonical models

Model	MAE (kg ha^{-1})	RMSE (kg ha^{-1})	R^2	MAPE (%)	Mean bias (kg ha^{-1})
Elastic Net	368.498	599.288	0.532	12.106	+39.862
Prior two-year moving average	382.996	609.323	0.516	12.161	-39.064
LSTM	400.087	602.435	0.527	13.023	+5.965

FT-Transformer	405.125	606.417	0.521	12.753	-32.078
SimpleRNN	406.841	602.509	0.527	13.168	+28.391

These results represent the one-time locked-test evaluation. No model, feature scenario, preprocessing rule, statistical procedure, or hyperparameter was modified in response to final-test performance.

4.3.2 Relative Final-Test Improvement

Relative to the prior two-year moving-average baseline, Elastic Net reduced MAE by $14.498\ \text{kgha}^{-1}$, corresponding to a 3.79% improvement. Its MAE was also $31.589\ \text{kgha}^{-1}$ lower than LSTM, $36.627\ \text{kgha}^{-1}$ lower than FT-Transformer, and $38.343\ \text{kgha}^{-1}$ lower than SimpleRNN, equivalent to reductions of 7.90%, 9.04%, and 9.42%, respectively.

The final-test ranking therefore differed from both development validation protocols: SimpleRNN had recorded the lowest district-grouped MAE, FT-Transformer the lowest rolling-origin MAE, and Elastic Net the lowest locked-test MAE. The final-test comparisons are descriptive and were not used for further model selection or tuning.

4.3.3 Year-Specific Final-Test Results

Year-specific performance varied substantially across the locked 2020-2022 period. In 2020, the prior two-year moving-average baseline achieved the lowest MAE (486.375 kg ha $^{-1}$), followed by SimpleRNN (512.470), Elastic Net (517.641), FT-Transformer (546.521), and LSTM ($551.118\ \text{kgha}^{-1}$). All models showed negative mean bias in 2020.

In 2021, Elastic Net achieved the lowest MAE ($398.534\ \text{kgha}^{-1}$), followed by SimpleRNN (402.574), LSTM (422.130), the prior two-year baseline (427.044), and FT-Transformer ($432.345\ \text{kgha}^{-1}$). All models showed positive mean bias. In 2022, Elastic Net again ranked first with an MAE of $189.320\ \text{kgha}^{-1}$ and the highest R^2 of 0.846, followed by LSTM (227.011), the prior two-year baseline (235.568), FT-Transformer (236.510), and SimpleRNN ($305.478\ \text{kgha}^{-1}$). Table 13 presents the complete year-specific final-test results.

Table 13. Year-specific performance of the five canonical models on the locked final test

Crop-year start	**Model**	**MAE (kg ha^{-1})**	**RMSE (kg ha^{-1})**	**R^2**	**Mean bias (kg ha^{-1})**
2020	Prior two-year moving average	486.375	694.703	0.456	-205.361

2020	Elastic Net	517.641	695.284	0.455	-69.673
2020	FT-Transformer	546.521	724.631	0.408	-107.298
2020	LSTM	551.118	721.078	0.414	-72.874
2020	SimpleRNN	512.470	708.080	0.435	-171.068
2021	Prior two-year moving average	427.044	701.916	0.420	+97.480
2021	Elastic Net	398.534	712.666	0.402	+133.925
2021	FT-Transformer	432.345	678.157	0.459	+31.589
2021	LSTM	422.130	684.575	0.448	+63.856
2021	SimpleRNN	402.574	671.807	0.469	+61.750
2022	Prior two-year moving average	235.568	372.188	0.753	-9.311
2022	Elastic Net	189.320	293.472	0.846	+55.335
2022	FT-Transformer	236.510	343.860	0.789	-20.526
2022	LSTM	227.011	316.521	0.821	+26.913
2022	SimpleRNN	305.478	369.253	0.756	+194.491

The annual rankings were therefore not fully stable. The prior two-year moving-average baseline ranked first in 2020, whereas Elastic Net ranked first in 2021 and 2022. The substantial decline in MAE across all models by 2022 indicates that final-test difficulty depended strongly on crop year.

4.3.4 Prediction versus Observed

Observed-versus-predicted patterns showed positive agreement across all three evaluation settings, although dispersion around the 1:1 line varied by protocol and yield range.

District-grouped validation showed comparatively tighter agreement, while rolling-origin validation exhibited greater dispersion, particularly at higher yields. The locked final test also showed substantial overlap among the five canonical models, with larger errors towards the extremes of the yield distribution.

Across panels, high-yield observations were often predicted below the identity line, indicating upper-range compression and underprediction of some high-yield cases. This pattern is examined further in the yield-range and outlier analyses. Figure 6 summarises observed-versus-predicted relationships for the five canonical models under district-grouped, rolling-origin, and locked final-test evaluation; the dashed 1:1 line indicates perfect agreement.

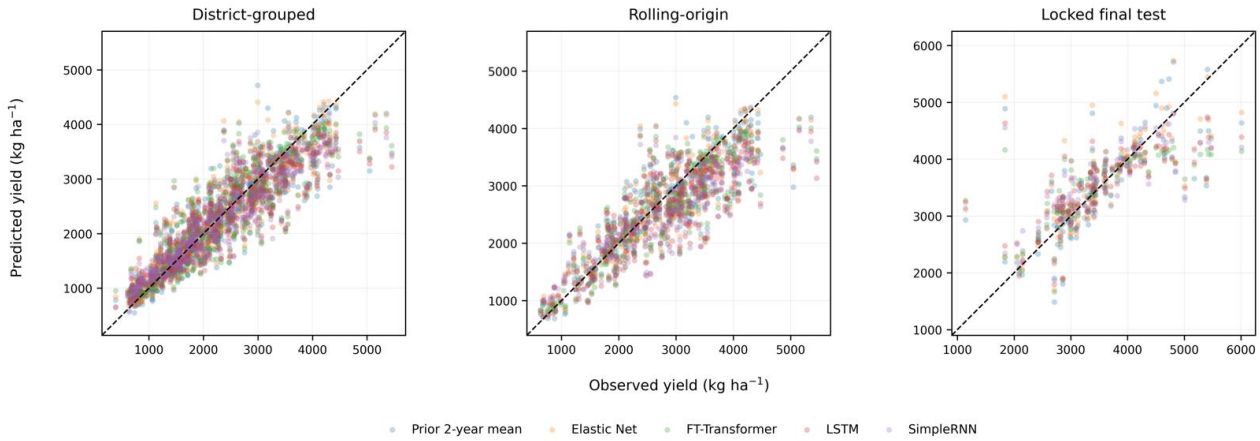


Figure 6. Observed versus predicted wheat yield across the three evaluation protocols.

4.4 Explainability Results

4.4.1 Tree-Based SHAP Analysis and Feature Interactions

Tree SHAP was applied to the development-selected history-only Extra Trees model using 114 primary chronological-validation observations without retraining or locked final-test data. The additivity audit confirmed numerical agreement between model predictions and reconstructed SHAP contributions.

The three-year prior mean yield was the dominant predictor, with a mean absolute SHAP value of 473.918 kg ha^{-1}, followed by the two-year prior mean (431.990) and one-year lagged yield (271.374 kg ha^{-1}). Three- and two-year prior yield standard deviations were much lower at 31.922 and 17.419 kg ha^{-1}. Panel A of Table 14 summarises these results. Figure 7 shows the clear dominance of recent productivity history over variability, coefficient-of-variation, and trend features.

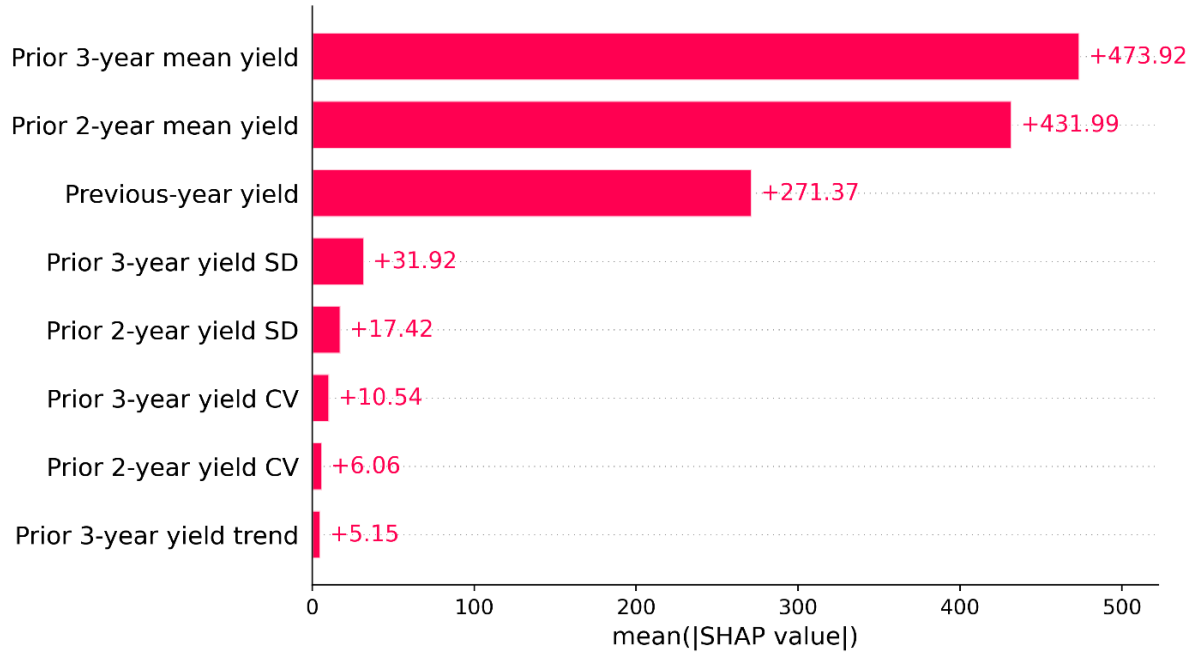


Figure 7. Global Tree SHAP importance of historical-yield predictors.

Figure 8 shows the observation-level SHAP distributions. Higher three- and two-year prior mean yields generally contributed positively to predictions, while lower values contributed negatively; previous-year yield showed a similar but weaker pattern.

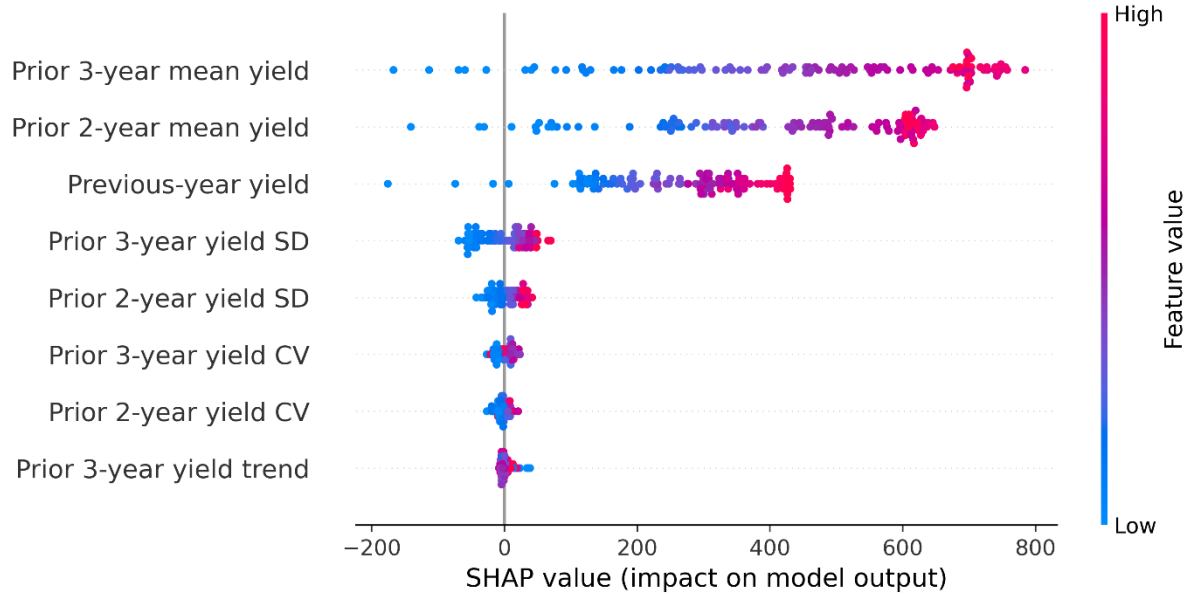


Figure 8. Tree SHAP summary for the historical-only Extra Trees model.

Development-only SHAP interaction analysis identified the strongest interaction between mean relative-humidity anomaly and rainy-day frequency (8.59). Rainy-day frequency also interacted strongly with heat-stress anomaly (8.14), solar-radiation anomaly (8.13), and mean wind speed (8.05), while relative-humidity anomaly × wind speed reached 8.04. Figure 9

summarises these interactions, which represent fitted-model joint contributions rather than causal meteorological effects.

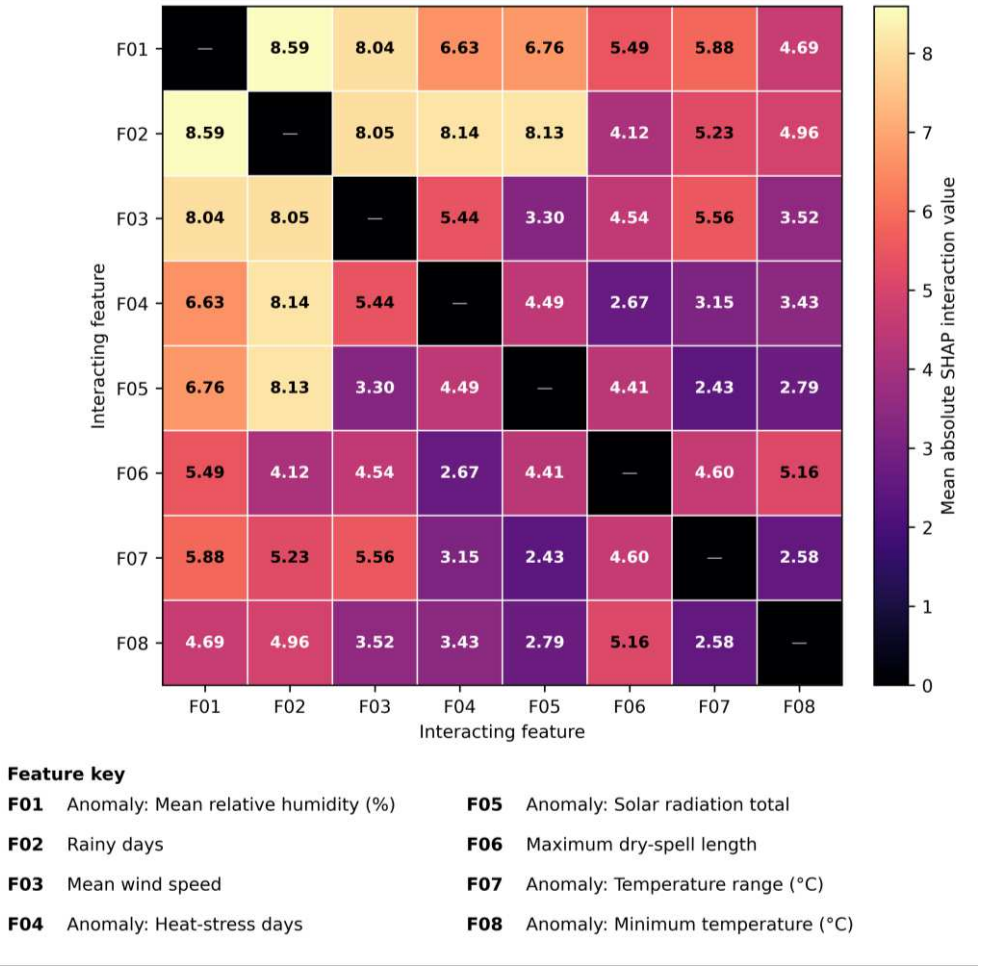


Figure 9. Development-only SHAP interaction strength among selected weather predictors.

The development-feature correlation network showed strong associations among temperature, temperature-anomaly, and growing-degree-day variables, with precipitation-related variables forming a separate structure. Figure 10 presents this network. Correlation reflects association among predictors, whereas SHAP interaction reflects joint contribution within the fitted model; neither establishes causality.

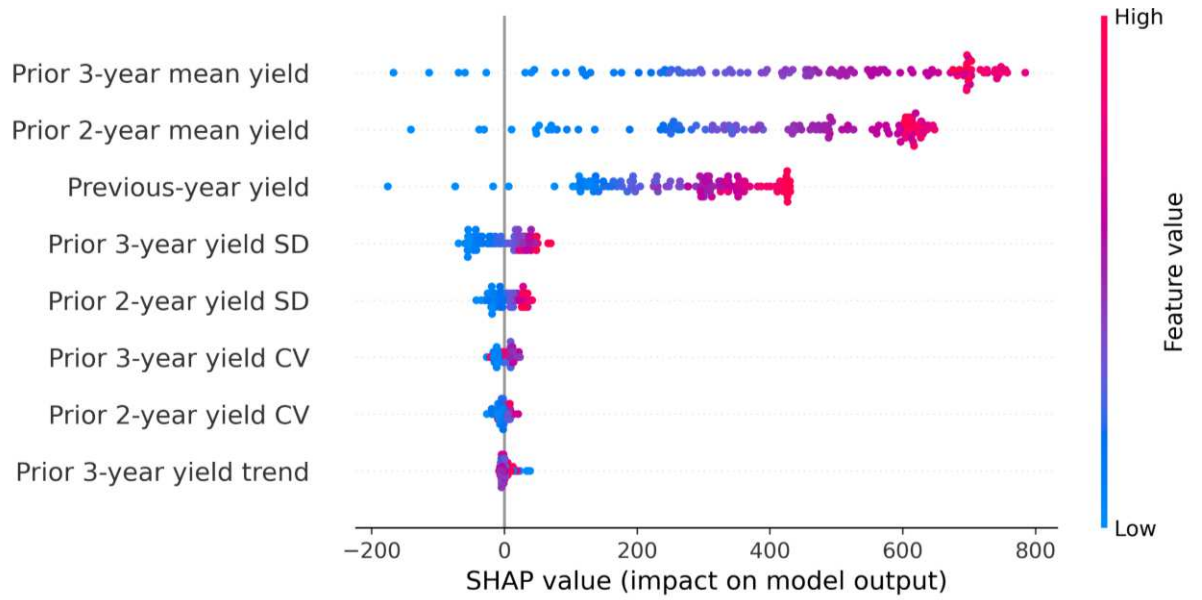


Figure 10. Development-feature correlation network.

4.4.2 Recurrent-Model Integrated Gradients

Integrated Gradients was applied to development-stage LSTM and SimpleRNN models using validated 64-point Gauss–Legendre integration across all 84 recurrent checkpoints. Attributions were assessed for static historical predictors, meteorological channels, month-specific inputs, and the seven October–April time steps.

Relative humidity was the highest-ranked meteorological channel for both models under both validation protocols. Under district-grouped validation, mean absolute attribution was 686.290 kg ha^{-1} for LSTM and 704.361 kg ha^{-1} for SimpleRNN. Solar radiation (394.133) and heat-stress days (272.933 kg ha^{-1}) followed for LSTM, while wind speed (451.203) and rainy days (240.130 kg ha^{-1}) followed for SimpleRNN.

Under rolling-origin validation, relative humidity remained dominant at 482.359 kg ha^{-1} for LSTM and 486.101 kg ha^{-1} for SimpleRNN. Rainy days (224.420) and solar radiation (214.274 kg ha^{-1}) followed for LSTM, whereas wind speed (342.864) and rainy days (292.557 kg ha^{-1}) followed for SimpleRNN. Panel B of Table 14 summarises these results.

April was the highest-ranked time step for both models under both protocols, although channel rankings were more protocol dependent. The attribution audit reproduced canonical predictions to numerical precision; the global 99th-percentile relative completeness error was 4.342%, and all 84 recurrent checkpoints satisfied the acceptance criteria.

Table 14. Principal explainability results

Panel A. Tree SHAP importance

Rank	Historical predictor	Mean absolute SHAP value (kg ha^{-1})
1	Three-year prior mean yield	473.918
2	Two-year prior mean yield	431.990
3	One-year lagged yield	271.374
4	Three-year prior yield standard deviation	31.922
5	Two-year prior yield standard deviation	17.419

Panel B. Leading recurrent weather-channel attributions

Protocol	Model	Weather channel	Mean absolute attribution (kg ha^{-1})
District grouped	LSTM	Relative humidity	686.290
District grouped	LSTM	Solar radiation	394.133
District grouped	LSTM	Heat-stress days	272.933
District grouped	SimpleRNN	Relative humidity	704.361
District grouped	SimpleRNN	Wind speed	451.203
District grouped	SimpleRNN	Rainy days	240.130
Rolling origin	LSTM	Relative humidity	482.359
Rolling origin	LSTM	Rainy days	224.420
Rolling origin	LSTM	Solar radiation	214.274
Rolling origin	SimpleRNN	Relative humidity	486.101
Rolling origin	SimpleRNN	Wind speed	342.864
Rolling origin	SimpleRNN	Rainy days	292.557

4.5 Explanation Stability and Cross-Method Agreement

Explanation stability was assessed using development-only analyses. For Elastic Net, exact linear SHAP values were evaluated across 1,000 district-cluster bootstrap resamples of the 608-observation development panel. All eight historical-yield predictors retained the same rank. The prior two-year mean yield remained first, with a mean absolute SHAP value of 362.367 kg ha^{-1} and a 95% bootstrap interval of 335.866–389.189 kg ha^{-1}.

For recurrent models, Integrated-Gradients rankings were compared with frozen-model neutralisation rankings. Under district-grouped validation, LSTM showed Spearman correlations of 0.857 for monthly time steps and 0.846 for weather channels; SimpleRNN

showed 0.464 and 0.791, respectively. Top-three overlap was two of three time steps for both models, three of three weather channels for LSTM, and two of three for SimpleRNN.

Agreement weakened under rolling-origin validation. Time-step correlations were −0.500 for LSTM and −0.357 for SimpleRNN, with top-three overlaps of one of three and zero of three. Weather-channel correlations were 0.033 and 0.412, with overlaps of zero of three and two of three, respectively. Table 15 summarises the bootstrap-stability and cross-method agreement results.

Table 15. Explanation stability and cross-method agreement under development validation.

Analysis	**Protocol/model**	**Comparison**	**Spearman (\rho)**	**Top-three overlap**	**Principal result**
Linear SHAP bootstrap	Elastic Net	Eight historical predictors	—	—	All eight ranks unchanged across 1,000 district-cluster bootstrap resamples; prior two-year mean remained rank 1
IG vs neutralisation	District grouped, LSTM	Monthly time steps	0.857	2/3	Strong agreement
IG vs neutralisation	District grouped, LSTM	Weather channels	0.846	3/3	Strong agreement
IG vs neutralisation	District grouped, SimpleRNN	Monthly time steps	0.464	2/3	Moderate agreement
IG vs neutralisation	District grouped, SimpleRNN	Weather channels	0.791	2/3	Strong agreement
IG vs neutralisation	Rolling origin, LSTM	Monthly time steps	−0.500	1/3	Weak/inconsistent agreement

IG vs neutralisation	Rolling origin, LSTM	Weather channels	0.033	0/3	Little agreement
IG vs neutralisation	Rolling origin, SimpleRNN	Monthly time steps	−0.357	0/3	Weak/inconsistent agreement
IG vs neutralisation	Rolling origin, SimpleRNN	Weather channels	0.412	2/3	Moderate agreement

Note: Elastic Net stability was evaluated using exact linear SHAP values. The prior two-year mean yield had a mean absolute SHAP value of 362.367 kg ha $^{-1}$*, with a 95% district-cluster bootstrap interval of 335.866–389.189* kg ha $^{-1}$*. Recurrent agreement compares Integrated-Gradients rankings with frozen-model neutralisation rankings.*

Overall, historical-feature importance was highly stable for Elastic Net, whereas recurrent meteorological and temporal rankings were more sensitive to the validation protocol. Integrated Gradients and frozen-model neutralisation therefore provided complementary evidence of fitted-model dependence rather than interchangeable measures of feature importance. All attribution and perturbation results were interpreted as properties of the fitted predictive models and not as causal effects of historical productivity or meteorological conditions.

4.6 Historical-Yield Contribution and Frozen-Model Perturbation Results

4.6.1 Matched LSTM Historical-Feature Ablation

A matched LSTM ablation quantified the incremental contribution of historical-yield information beyond the monthly meteorological sequence. Two configurations were compared: a weather-only LSTM and a weather-plus-history LSTM using the same seven-month weather sequence plus eight static historical-yield predictors. Recurrent architecture, weather inputs, optimisation settings, validation assignments, random seeds, and eligible observations were otherwise held constant. The experiment comprised two feature configurations, three seeds, five district-grouped folds, and nine rolling-origin folds, yielding 84 completed runs and 5,700 prediction records.

Under district-grouped validation, MAE decreased from 596.727 to 297.939 kgha^{-1} after adding historical predictors, a reduction of 298.788 kgha^{-1}(50.1%). The combined model produced lower absolute error for 73.19% of observations. The 95% confidence interval for the weather-plus-history minus weather-only difference was -403.049 to -209.403 kg ha^{-1}, and the district-cluster sign-flip comparison remained significant after Holm correction ($p < 0.001$).

Under rolling-origin validation, MAE decreased from 1,067.055 to $433.503\ \mathrm{kgha}^{-1}$, a reduction of $633.551\ \mathrm{kgha}^{-1}(59.4\%)$, with the combined model outperforming the weather-only model for 83.63% of observations. The 95% confidence interval was -893.283 to -392.265 kg ha $^{-1}$, and the exact crop-year sign-flip test remained significant after Holm correction ($p = 0.0039$). Panel A of Table 16 summarises the matched ablation results.

These results show that historical productivity added substantial predictive value beyond the monthly meteorological sequence under both validation protocols. The ablation is a controlled feature-block comparison and should not be substituted for the canonical LSTM benchmark results.

4.6.2 Execution and Applicability Audit

The frozen-model perturbation analysis evaluated four active fitted models-Elastic Net, FT-Transformer, LSTM, and SimpleRNN-under district-grouped and rolling-origin validation. The prior two-year moving-average baseline was retained as a reference-only model because its prediction rule does not contain a fitted feature representation suitable for neutralisation.

The analysis comprised 44 model-specific perturbation definitions evaluated under both validation protocols, yielding 88 model-protocol metric records, 41,800 ensemble predictions, and 122,550 seed-level deep-model predictions. Unmodified model replay reproduced the archived canonical predictions to numerical precision, with a maximum discrepancy of $9.095 \times 10^{-13}\ \mathrm{kgha}^{-1}$. No model was retrained, and no locked final-test prediction or metric was generated.

The applicability audit considered complete historical-feature blocks, seasonal raw-weather and weather-anomaly blocks, individual recurrent weather channels, and monthly time steps. Historical-feature neutralisation was applicable to all four active models. Elastic Net and FT-Transformer used historical-only inputs, whereas LSTM and SimpleRNN combined the seven-month meteorological sequence with eight static historical predictors. Consequently, only the recurrent models were eligible for individual weather-channel and monthly time-step perturbations. No active canonical model used a separate seasonal raw-weather or weather-anomaly block. The executed perturbation set therefore comprised four historical-block definitions, 26 recurrent-channel definitions, and 14 recurrent time-step definitions across the two validation protocols.

4.6.3 Historical-Feature Neutralisation

Neutralisation of the complete historical-yield block produced the largest and most consistent deterioration among the perturbation families. Under district-grouped validation, Elastic Net

MAE increased from 311.687 to 842.725 kgha^{-1}, an increase of 531.038 kgha^{-1} or 170.4%. FT-Transformer increased from 320.752 to 835.541 kgha^{-1}, an increase of 514.789 kgha^{-1} or 160.5%. Its RMSE increased from 438.043 to 1,004.549 kgha^{-1}, while R^2 decreased from 0.808 to -0.007.

The recurrent models showed the same dependence on historical productivity information. LSTM MAE increased from 307.407 to 737.173 kgha^{-1}, an increase of 429.766 kgha^{-1} or 139.8%, while SimpleRNN increased from 304.208 to 742.891 kgha^{-1}, an increase of 438.683 kgha^{-1} or 144.2%. For LSTM, RMSE increased from 415.320 to 893.772 kgha^{-1} and R^2 decreased from 0.828 to 0.203. Historical-block neutralisation significantly degraded district-grouped accuracy for all four active models after Holm correction.

The same directional effect occurred under rolling-origin validation. Elastic Net MAE increased from 389.087 to 1,044.201 kgha^{-1}, a difference of 655.114 kgha^{-1} or 168.4%. The corresponding increases were 624.361 kgha^{-1} for FT-Transformer, 546.955 kgha^{-1} for LSTM, and 502.624 kgha^{-1} for SimpleRNN. Although all four models deteriorated substantially following historical-feature neutralisation, none of the rolling-origin contrasts remained statistically resolved after Holm correction across the nine crop-year clusters. These results are summarised in Panel B of Table 16.

Table 16. Matched historical-feature ablation and principal frozen-model perturbation results

Panel A. Matched LSTM historical-feature ablation.

Validation protocol	**Weather-only MAE (kgha^{-1})**	**Weather + history MAE (kgha^{-1})**	**Difference: combined − weather-only (kgha^{-1})**	**Relative change (%)**	**95% CI (kgha^{-1})**	**Combined-model win rate (%)**	**Holm-adjusted (p)**
District grouped	596.727	297.939	−298.788	−50.1	−403.049 to −209.403	73.19	<0.001

Rolling origin	1,067.055	433.503	−633.551	−59.4	−893.283 to −392.265	83.63	0.0039

Panel B. Historical-yield-block neutralisation

Protocol	**Model**	**Reference MAE ($\mathbf{kgha^{-1}}$)**	**Perturbed MAE ($\mathbf{kgha^{-1}}$)**	**MAE increase ($\mathbf{kgha^{-1}}$)**	**Relative increase (%)**	**95% CI for MAE increase ($\mathbf{kgha^{-1}}$)**	**Holm-adjusted (p)**	**Statistical outcome**
District grouped	Elastic Net	311.687	842.725	531.038	170.4	462.356–602.426	0.0008	Significant
District grouped	FT-Transformer	320.752	835.541	514.789	160.5	451.083–581.103	0.0008	Significant
District grouped	LSTM	307.407	737.173	429.766	139.8	368.767–494.071	0.0008	Significant
District grouped	SimpleRNN	304.208	742.891	438.683	144.2	371.652–506.356	0.0008	Significant

Rolling origin	Elastic Net	389.087	1,044.201	655.114	168.4	484.145–812.552	0.28125	Not statistically resolved
Rolling origin	FT-Transformer	367.610	991.970	624.361	169.8	464.915–777.348	0.28125	Not statistically resolved
Rolling origin	LSTM	430.873	977.828	546.955	126.9	402.431–691.183	0.28125	Not statistically resolved
Rolling origin	SimpleRNN	459.834	962.458	502.624	109.3	383.105–618.893	0.28125	Not statistically resolved

4.6.4 Weather-Channel and Monthly Time-Step Sensitivity

Weather-channel neutralisation produced substantially smaller MAE changes than complete historical-block neutralisation and showed greater variation across models and validation protocols. Mean relative humidity provided the strongest cross-model channel signal: neutralisation increased MAE in all four recurrent model-protocol contexts, with a mean increase of $53.155\ \text{kgha}^{-1}$, equivalent to 15.57%. Under district-grouped validation, relative-humidity neutralisation significantly degraded both LSTM and SimpleRNN performance after Holm correction.

Heat-stress days ranked second across the combined channel analysis, with a mean MAE increase of $30.435\ \text{kgha}^{-1}$ or 7.90%, with deterioration in all four contexts. Its district-grouped effect remained statistically significant for LSTM. Mean wind speed ranked third, with a mean increase of $17.659\ \text{kgha}^{-1}$ or 4.60%, although no individual wind-speed contrast remained significant after multiplicity correction. These leading cross-model effects are summarised in Panel C of Table 16.

Channel rankings also differed between validation protocols. Under district-grouped validation, relative humidity ranked first for both recurrent models, whereas under rolling-origin validation heat-stress days ranked first for LSTM and relative humidity ranked first for

SimpleRNN. No rolling-origin weather-channel perturbation remained statistically significant after Holm correction, indicating weaker statistical resolution of individual meteorological dependencies under future-year evaluation.

Panel C. Leading cross-model weather-channel neutralisation effects

Weather channel	**Contexts with increased MAE**	**Mean MAE increase ($kgha^{-1}$)**	**Range of MAE increase ($kgha^{-1}$)**	**Mean relative increase (%)**	**Consensus rank**
Mean relative humidity	4/4	53.155	9.641–78.445	15.57	1
Heat-stress days	4/4	30.435	8.094–61.164	7.90	2
Mean wind speed	4/4	17.659	9.893–32.574	4.60	3

At the monthly time-step level, the strongest statistically resolved effects occurred under district-grouped validation. Neutralising April increased LSTM MAE from 307.407 to 347.580 $kgha^{-1}$, an increase of 40.173 $kg\ ha^{-1}$. Neutralising March increased LSTM MAE to 335.431$kg\ ha^{-1}$, corresponding to an increase of 28.024 $kgha^{-1}$. Both effects remained significant after Holm correction.

For SimpleRNN, March neutralisation increased MAE by 25.220 $kgha^{-1}$, February by 21.059 $kgha^{-1}$, and April by 18.792 $kgha^{-1}$. All three effects remained significant after multiplicity adjustment. Across the combined model-protocol analysis, April had the highest overall time-step consensus rank, although the ordering of individual months varied between models and protocols. No rolling-origin monthly time-step comparison remained statistically significant after Holm correction. The statistically resolved district-grouped effects are reported in Panel D of Table 16.

Panel D. Statistically resolved district-grouped monthly time-step effects

Model	Neutralised month	Reference MAE (kg ha^{-1})	Perturbed MAE (kg ha^{-1})	MAE increase (kg ha^{-1})	95% CI (kg ha^{-1})	Holmadjusted p
LSTM	April	307.407	347.580	40.173	24.269-57.978	0.00164
LSTM	March	307.407	335.431	28.024	14.539-40.960	0.02212
SimpleRNN	March	304.208	329.428	25.220	11.132-38.777	0.01360
SimpleRNN	February	304.208	325.267	21.059	12.008-30.023	0.00729
SimpleRNN	April	304.208	323.000	18.792	10.570-27.064	0.01360

Taken together, the matched ablation and frozen-model perturbation analyses showed that historical productivity provided the dominant predictive information within the available district-level feature set. Adding historical predictors to an otherwise matched recurrent weather model reduced MAE by 50.1% under district-grouped validation and 59.4% under rolling-origin validation, while neutralising the historical block in already fitted models produced substantially larger performance losses than neutralising individual weather channels

or monthly positions. Meteorological variables nevertheless retained measurable predictive contributions, particularly through relative humidity, heat-stress exposure, and late-season time steps. These findings therefore indicate a strong incremental contribution from historical productivity without implying that meteorological information was universally unnecessary.

4.7 Effective Sequence-Length Sensitivity

The sequence-length analysis evaluated 84 recurrent checkpoints across 14 development folds. It generated 39,900 seed-level predictions and 13,300 ensemble predictions without retraining either recurrent model. Full seven-month replay reproduced the archived raw and scaled reference predictions with zero error. No locked final-test data were accessed.

Under district-grouped validation, restricting the effective LSTM sequence to one, two, three, or four months increased MAE by 35.991, 52.938, 47.579, and $20.513\ \mathrm{kgha}^{-1}$, respectively. All four differences remained significant after Holm correction. The five- and six-month configurations differed from the seven-month result by only 2.029 and $-0.078\ \mathrm{kgha}^{-1}$, respectively, and neither difference was statistically resolved.

For SimpleRNN, retaining one, two, three, or four months increased MAE by 45.494, 80.456, 103.122, and $42.222\ \mathrm{kgha}^{-1}$, respectively. All four differences remained significant after Holm correction. The five- and six-month configurations increased MAE by 9.223 and $4.230\ \mathrm{kgha}^{-1}$, respectively, but neither remained statistically significant after adjustment.

Under rolling-origin validation, the complete seven-month sequence produced the lowest MAE for both recurrent models. LSTM recorded its highest MAE with two retained months, whereas SimpleRNN recorded its highest MAE with three retained months. Although every shortened rolling-origin sequence produced a higher MAE than the corresponding seven-month configuration, none of these differences remained statistically significant after Holm correction. The complete sequence-length results are reported in Table 17.

Table 17. Recurrent-model MAE by effective sequence length.

Retained months	District-grouped LSTM	District-grouped SimpleRNN	Rolling-origin LSTM	Rolling-origin SimpleRNN
1	343.398*	349.702*	471.258	487.599
2	360.345*	384.664*	504.652	557.767
3	354.986*	407.329*	499.409	610.708
4	327.920*	346.429*	466.085	536.048

5	309.436	313.431	443.952	510.319
6	307.329	308.438	434.309	496.831
7	307.407	304.208	430.873	459.834

*Significantly higher than the corresponding seven-month result after Holm correction. Table entries are MAE in kg ha^{-1}.

The district-grouped sensitivity profiles show that both recurrent models incurred substantially greater error when the effective sequence was restricted to one through four months. Performance approached the complete-sequence result after five months for LSTM and after six months for SimpleRNN, while the seven-month points reproduce the canonical district-grouped MAEs. Figure 11 presents these relationships.

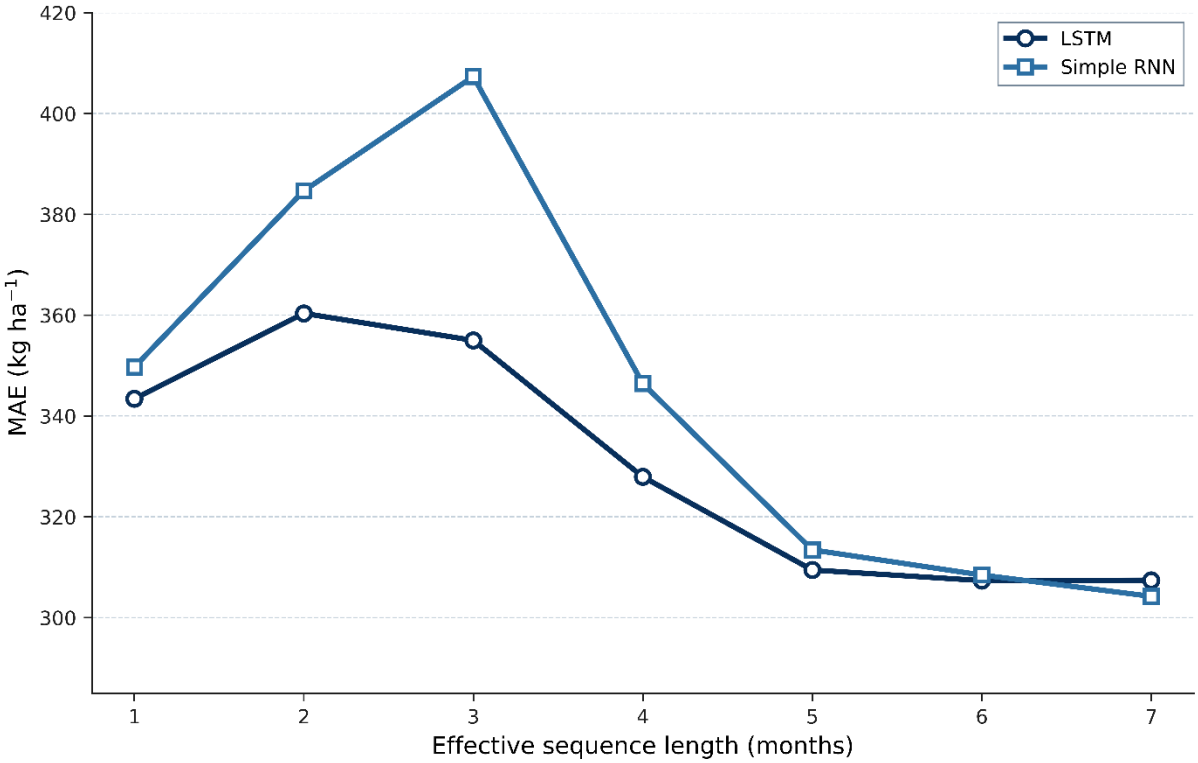


Figure 11. Effective sequence-length sensitivity under district-grouped validation.

The rolling-origin profiles showed a broadly similar reduction in error as additional crop-season months were restored. LSTM reached its highest MAE with two retained months, whereas the SimpleRNN maximum occurred with three months. For both models, the full October–April sequence produced the lowest rolling-origin MAE, although none of the shortened-sequence comparisons remained statistically significant after Holm correction. Figure 12 presents the corresponding temporal sensitivity results.

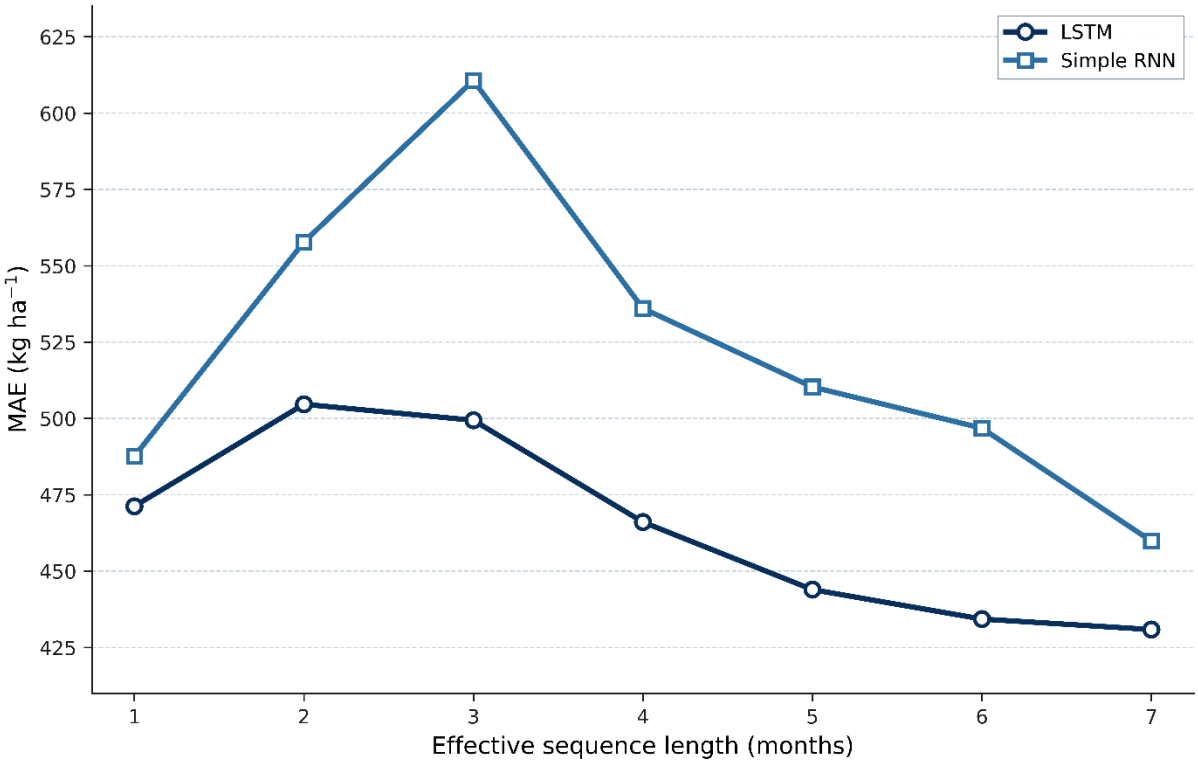


Figure 12. Effective sequence-length sensitivity under rolling-origin validation.

4.8 When Models Fail: Error and Failure Analysis

4.8.1 Residual Distributions

Residual distributions were examined across district-grouped validation, rolling-origin validation, and the locked final test. Residuals were defined as predicted minus observed yield; negative values therefore indicate underprediction.

Under district-grouped validation, mean bias remained comparatively close to zero: -47.370 kgha^{-1} for the prior two-year moving-average baseline, +2.663 kg ha^{-1} for Elastic Net, -11.284 kg ha $^{-1}$ for FT-Transformer, -5.052 kgha^{-1} for LSTM, and -8.925 kgha^{-1} for SimpleRNN. Thus, pooled spatial-transfer errors showed limited systematic directional bias.

Rolling-origin validation exhibited a clearer negative shift. Mean bias was $-106.424, -163.354, -154.567$, -224.799, and -253.354 kg ha^{-1} for the prior two-year baseline, Elastic Net, FT-Transformer, LSTM, and SimpleRNN, respectively. The recurrent models therefore showed the strongest systematic underprediction, with residual distributions extending more strongly into the negative-error region.

On the locked final test, pooled bias again moved closer to zero: -39.064 kgha^{-1} for the prior two-year baseline, $+39.862$ kgha^{-1} for Elastic Net, -32.078 kg ha^{-1} for FT-Transformer, +5.965 kg ha^{-1} for LSTM, and $+28.391$ kgha^{-1} for SimpleRNN.

Overall, residual behaviour was protocol dependent: district-grouped errors were comparatively balanced, rolling-origin evaluation showed systematic underprediction, and the locked final test showed smaller model-specific directional shifts. The strongest negative errors occurred predominantly among higher-yield observations, consistent with the yield-range analysis in Section 4.8.4. These protocol-specific residual distributions are shown in Figure 13.

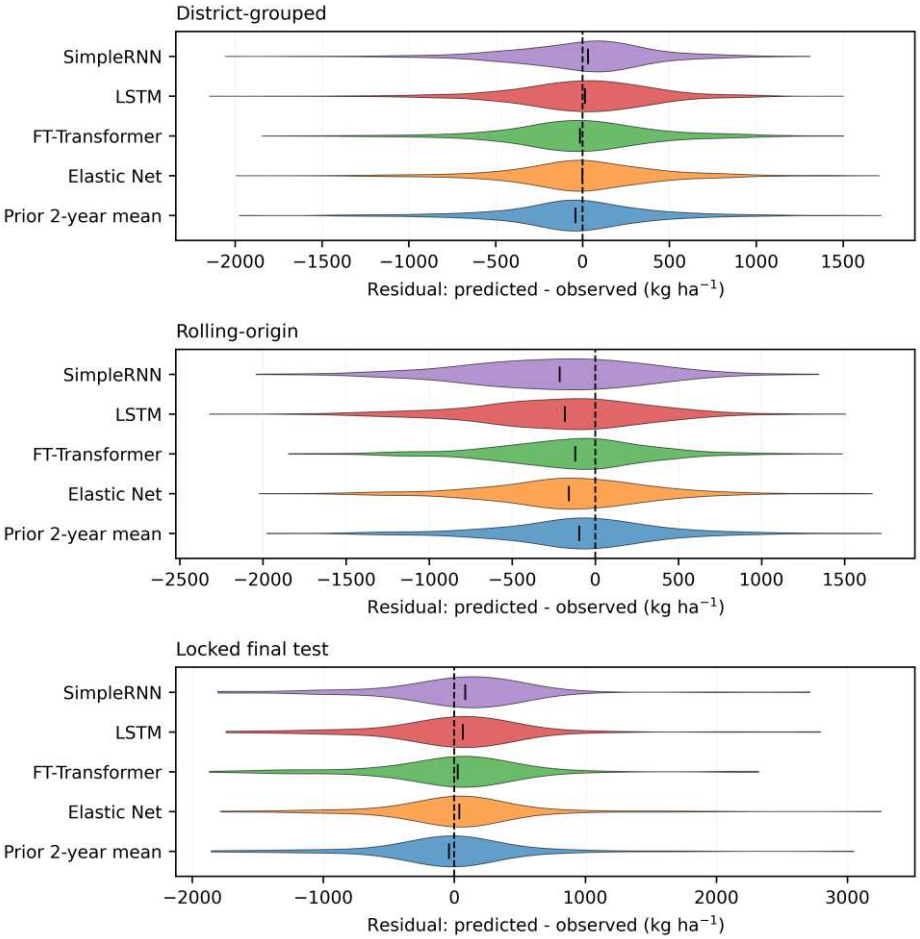


Figure 13. Residual distributions across the three evaluation protocols.

4.8.2 District-Level Error Patterns

Prediction difficulty also varied geographically. Under district-grouped validation, Ujjain was the highesterror district for all five canonical models. District-specific MAE was $631.845\ \text{kgha}^{-1}$ for the prior two-year baseline, $668.192\ \text{kgha}^{-1}$ for Elastic Net, $655.470\ \text{kgha}^{-1}$ for FT-Transformer, $578.620\ \text{kgha}^{-1}$ for LSTM, and $559.399\ \text{kgha}^{-1}$ for SimpleRNN. The recurrence of Ujjain across model families indicates that this spatial difficulty was not architecture specific.

The pattern changed under rolling-origin validation. Dewas had the highest district-specific MAE for the prior two-year baseline ($696.648\ \text{kgha}^{-1}$), Elastic Net ($703.850\ \text{kgha}^{-1}$), LSTM ($732.844\ \text{kgha}^{-1}$), and SimpleRNN ($757.264\ \text{kgha}^{-1}$). FT-Transformer differed,

recording its highest MAE in Hoshangabad at 639.759 kgha^{-1}. In contrast, the lowest-error locations varied across models and protocols rather than forming a common low-error region. These results demonstrate that spatial prediction difficulty depended on the type of generalisation being assessed. Figure 14 maps district-specific MAE for the numerically leading model under each evaluation protocol-SimpleRNN under district-grouped validation, FT-Transformer under rolling-origin validation, and Elastic Net on the locked final test-and illustrates the corresponding change in spatial error patterns.

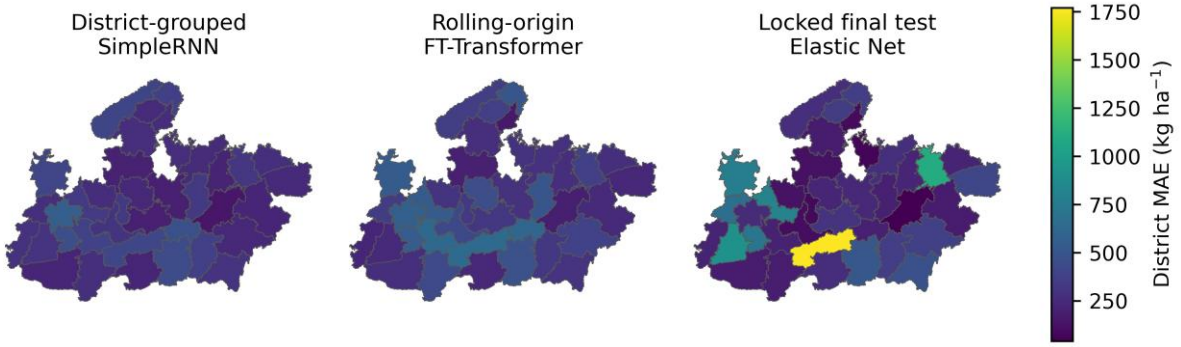


Figure 14. District-level MAE of the protocol-leading model under each evaluation protocol.

4.8.3 Temporal Drift

Prediction error also varied substantially across crop years. Under district-grouped validation, crop-year start 2019 produced the highest annual MAE for all five canonical models: 540.477 kgha^{-1} for the prior two-year baseline, 538.691 kgha^{-1} for Elastic Net, 537.904 kgha^{-1} for FT-Transformer, 607.705 kgha^{-1} for LSTM, and 598.038 kgha^{-1} for SimpleRNN.

Under rolling-origin validation, the prior two-year baseline, Elastic Net, and FT-Transformer again recorded their highest annual errors in 2019, with MAEs of 540.477, 545.196, and 530.738 kgha^{-1}, respectively. In contrast, the recurrent models were most strongly affected in 2014: LSTM reached 726.134 kgha^{-1} and SimpleRNN 761.094 kgha^{-1}, with mean biases of -711.999 and -750.839 kgha^{-1}, respectively.

The pooled temporal penalty was also substantial. Rolling-origin minus district-grouped MAE increased by 57.760 kgha^{-1} for the prior two-year baseline, 77.400 kgha^{-1} for Elastic Net, 46.858 kgha^{-1} for FT-Transformer, 123.466 kgha^{-1} for LSTM, and 155.626 kgha^{-1} for SimpleRNN.

On the locked final test, mean MAE across the five canonical models declined from 522.825 kgha^{-1} in 2020 to 416.526 kgha^{-1} in 2021 and 238.777 kgha^{-1} in 2022. The prior

two-year baseline was best in 2020 (486.375 kg ha^{-1}), whereas Elastic Net was best in 2021 (398.534 kg ha^{-1}) and 2022 (189.320 kg ha^{-1}). Temporal difficulty and model ranking were therefore not constant across crop years. Figure 15 summarises these temporal changes across development validation and the locked final-test period.

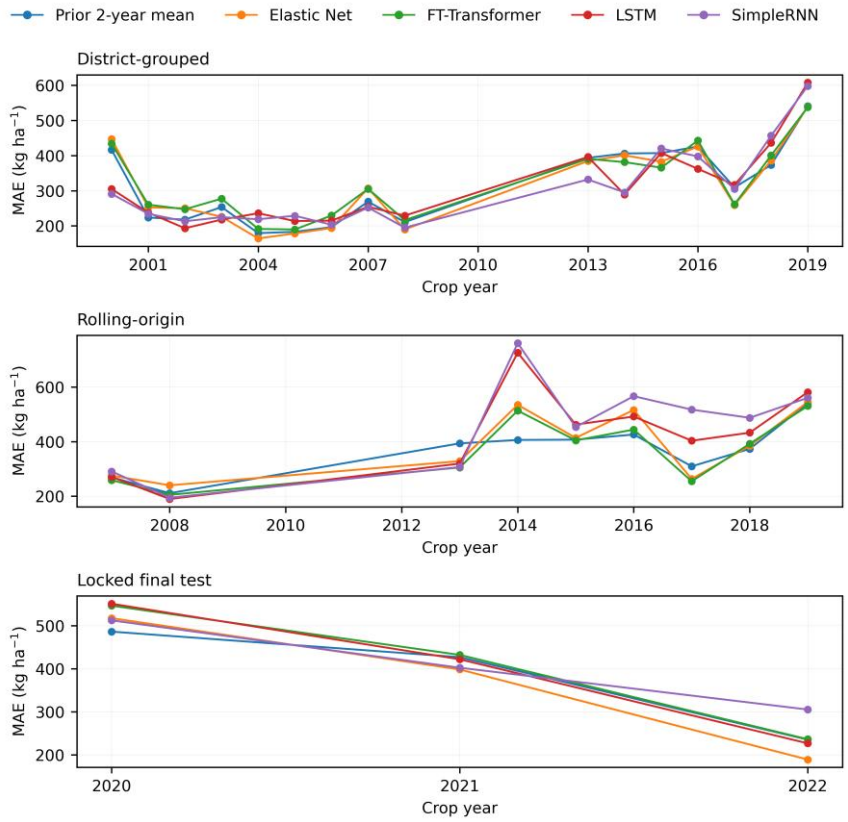


Figure 15. Temporal variation in model error across development validation and the locked final test.

4.8.4 Yield-Range and High-Yield Errors

The training-derived Tukey limits were approximately -495.45 and 4,576.92 kgha^{-1}. Because wheat yield was non-negative, only the upper limit was operationally relevant. Five high-yield observations exceeded this threshold under each development validation protocol and were retained in all principal metrics.

Under district-grouped validation, non-outlier MAE ranged from 294.257 to 311.343 kgha^{-1}, whereas high-yield-outlier MAE ranged from 1,370.448 to 1,637.947 kgha^{-1}. Under rolling-origin validation, the corresponding ranges were $352.084 - 443.798$ and $1{,}414.020 - 1{,}611.244$ kgha^{-1}. Depending on model and protocol, high-yield-outlier MAE was approximately 3.47-5.53 times the corresponding non-outlier MAE. This pronounced outlier sensitivity is shown in Figure 16.

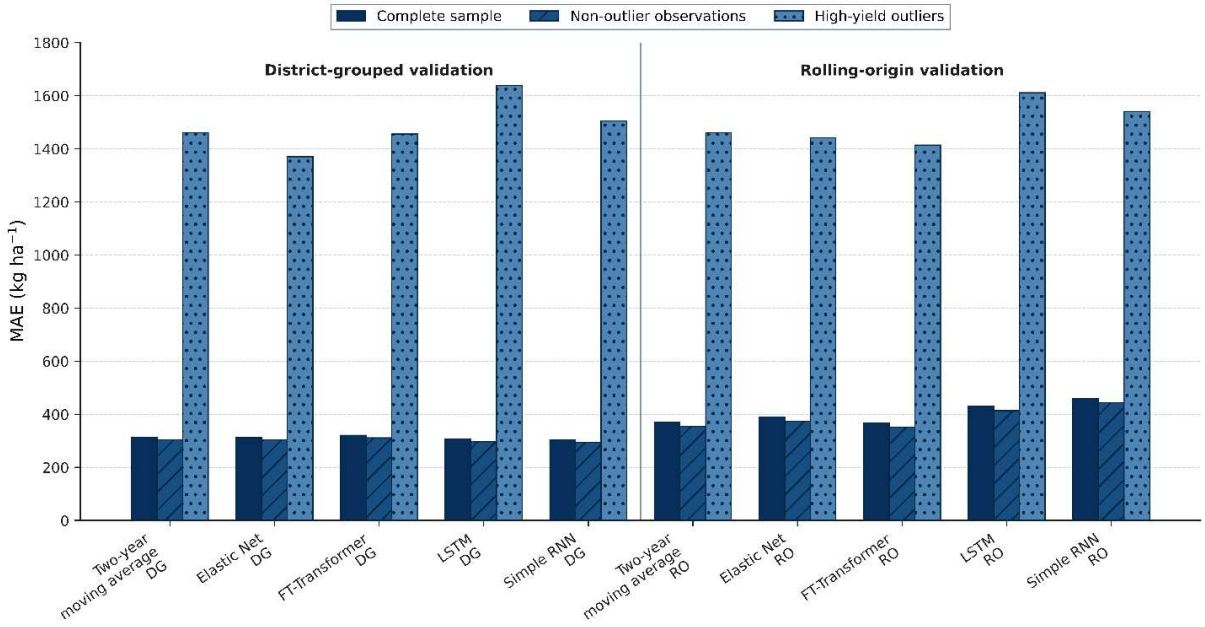


Figure 16. Model sensitivity to high-yield observations identified using the training-derived Tukey rule.

The broader yield-range analysis showed the same upper-range difficulty. District-grouped validation contained 61 lower-range, 486 central-range, and 61 upper-range observations, whereas rolling-origin validation contained 35, 272, and 35, respectively.

Under district-grouped validation, the prior two-year baseline achieved the lowest lower-range MAE (146.907 kgha^{-1}), SimpleRNN the lowest central-range MAE (278.603 kgha^{-1}), and Elastic Net the lowest upper-range MAE (499.238 kgha^{-1}). Under rolling-origin validation, FT-Transformer achieved the lowest lower- and central-range MAEs (198.350 and 345.016 kgha^{-1}), whereas the prior two-year baseline had the lowest upper-range MAE (608.531 kgha^{-1}). Nevertheless, the upper-yield range produced the largest MAE for every canonical model under both protocols. The complete lower-, central-, and upper-range error profiles are presented in Figure 17, while the corresponding outlier- and yield-range-specific MAE values are summarised in Table 18.

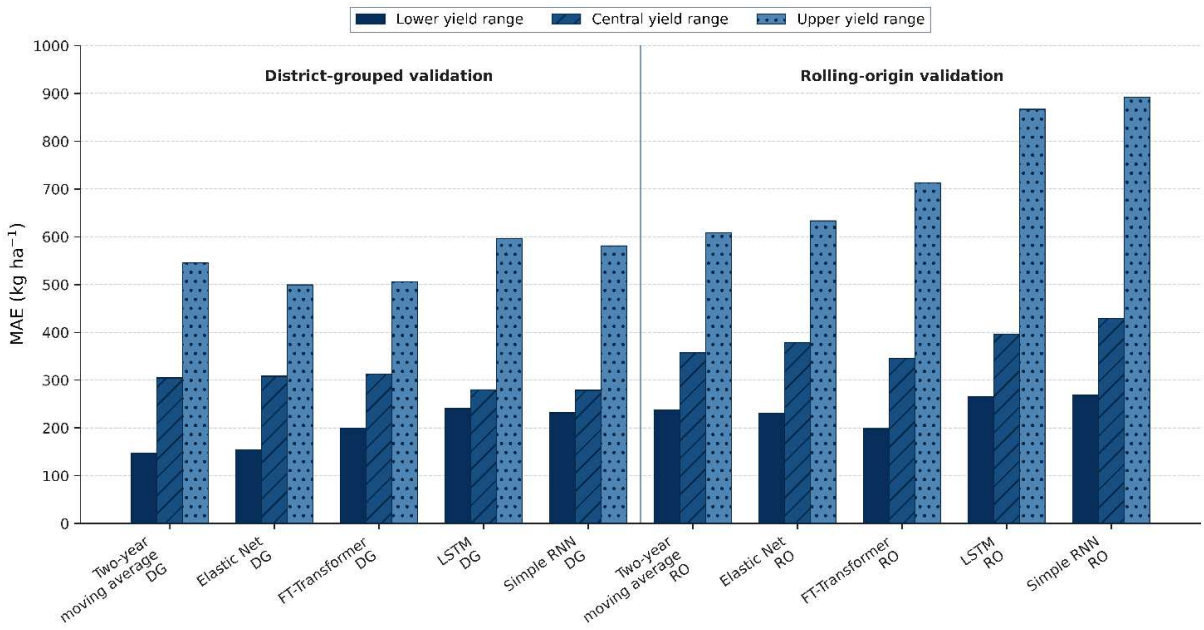


Figure 17. Model error across lower, central, and upper observed-yield ranges.

The consistent deterioration at the upper end of the yield distribution identifies high productivity as a reproducible failure region. This pattern is also consistent with the compression of high-yield predictions towards the central yield range observed in Figure 6.

Table 18. Outlier and yield-range sensitivity

Panel A. MAE by IQR outlier class

Protocol	Model	Complete sample	Non-outlier	High-yield outlier
District grouped	Prior two-year moving average	312.902	303.387	1,460.443
District grouped	Elastic Net	311.687	302.908	1,370.448
District grouped	FT-Transformer	320.752	311.343	1,455.506
District grouped	LSTM	307.407	296.374	1,637.947
District grouped	SimpleRNN	304.208	294.257	1,504.324

Rolling origin	Prior two-year moving average	370.662	354.493	1,460.443
Rolling origin	Elastic Net	389.087	373.471	1,441.600
Rolling origin	FT-Transformer	367.610	352.084	1,414.020
Rolling origin	LSTM	430.873	413.360	1,611.244
Rolling origin	SimpleRNN	459.834	443.798	1,540.638

Panel B. MAE by observed-yield range

Protocol	**Model**	**Lower range**	**Central range**	**Upper range**
District grouped	Prior two-year moving average	146.907	304.583	545.179
District grouped	Elastic Net	154.284	307.903	499.238
District grouped	FT-Transformer	199.478	312.732	505.921
District grouped	LSTM	240.446	279.617	595.782
District grouped	SimpleRNN	231.537	278.603	580.876
Rolling origin	Prior two-year moving average	237.756	357.155	608.531
Rolling origin	Elastic Net	230.296	378.101	633.255
Rolling origin	FT-Transformer	198.350	345.016	712.453
Rolling origin	LSTM	264.881	396.166	866.587
Rolling origin	SimpleRNN	268.157	428.919	891.768

Table entries are MAE in kgha^{-1}.

4.8.5 Bland-Altman Agreement

Bland-Altman analysis provided a complementary assessment of prediction agreement relative to the mean observed and predicted yield. For the protocol-leading models, SimpleRNN under district-grouped validation had a mean difference of $-8.9\ \text{kgha}^{-1}$, with approximate 95% limits of agreement from -816.0 to $+798.1\ \text{kgha}^{-1}$. FT-Transformer under rolling-origin validation had a mean difference of -154.6 kg ha-1 , with limits from $-1{,}083.8$ to

$+774.7$ kgha^{-1}. On the locked final test, Elastic Net had a mean difference of $+39.9$ kgha^{-1}, with limits from -1,137.3 to +1,217.0 kg ha^{-1}.

The broad limits of agreement demonstrate that small pooled mean biases did not imply uniformly close agreement at the individual district-crop-year level. Error dispersion also increased towards higher mean yield levels, consistent with the high-yield and upper-range analyses. Rolling-origin validation additionally showed a downward shift consistent with its systematic underprediction. These agreement patterns and limits are displayed in Figure 18.

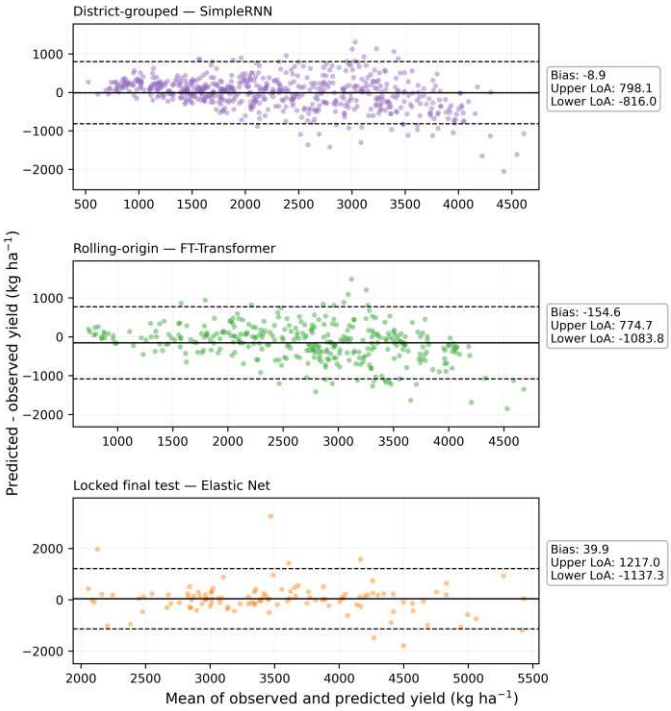


Figure 18. Bland-Altman agreement for the protocol-leading model under each evaluation protocol.

4.8.6 Weather-Extreme and Late-Sowing Failure Analysis

Development-stage predictions were additionally stratified using training-derived quartiles of seasonal rainfall, heat-stress exposure, and maximum dry-spell duration. These comparisons were descriptive and were not treated as event-labelled drought or heatwave analyses.

Under district-grouped validation, the highest-minus-lowest rainfall-quartile difference in MAE was small and model dependent, ranging from -15.740 to $+37.572$ kgha^{-1}. Under rolling-origin validation, however, all five canonical models showed higher MAE in the highest rainfall quartile, with increases of 72.200-151.854 kgha^{-1}. Thus, comparatively wet seasons were associated with greater prediction difficulty specifically under temporal generalisation.

The highest heat-stress quartile did not correspond to greater prediction error. Q4-minus-Q1 MAE differences were negative for every model under both protocols, ranging from -112.897

to -40.395 kgha^{-1}. Maximum dry-spell differences were likewise negative throughout, ranging from -181.944 to -23.553 kg ha $^{-1}$. The available evidence therefore does not support a claim that the highest heat-stress or longest dry-spell strata were systematically the most difficult to predict.

These conditional associations should not be interpreted causally because they may reflect concurrent differences in district, crop year, yield level, historical productivity, or unobserved management conditions. Feature-importance results were therefore not used as substitutes for explicit subgroup failure evidence.

Late-sowing failures could not be evaluated because sowing-date information was unavailable in the modelling dataset. Accordingly, no claim is made regarding prediction failure specifically attributable to delayed sowing. The failure analysis was therefore restricted to the executed residual, spatial, temporal, yield-range, outlier, agreement, and meteorological-stratification analyses.

4.9 Learning Behaviour, Model Capacity, and Reproducibility

4.9.1 Learning-Curve and Training-History Diagnostics

The learning-curve reporting package was generated entirely from completed development-stage validation records and saved neural optimisation histories, without retraining any model or accessing the locked final test. The package retained district-grouped and rolling-origin validation trajectories, epoch-level neural training histories, figure captions, source manifests, output metadata, and cryptographic checksums.

The rolling-origin trajectories represented progressively expanding training histories across successive validation years. Increasing the amount of preceding training data did not produce a monotonic reduction in validation error. In particular, the recurrent models exhibited pronounced increases in error in some later temporal folds despite the larger available training histories. This behaviour is consistent with the temporal-drift results in Section 4.8.3 and indicates that additional historical observations did not eliminate year-to-year distributional variation.

District-grouped validation showed a different form of variability because each fold withheld a different set of districts. Each fold was fitted independently, and the resulting variation was retained as part of the spatial-transfer diagnostic. Principal district-grouped performance was calculated from pooled out-of-fold predictions rather than from a simple average of fold-level MAEs.

The neural optimisation histories retained epoch-level training loss, inner-validation MAE, learning-rate changes, early-stopping status, and best-checkpoint information. Training loss generally declined during optimisation, whereas validation MAE typically reached a minimum and subsequently plateaued or fluctuated, providing the basis for early stopping and checkpoint restoration. Learning-rate reduction was triggered when validation improvement stalled according to the frozen scheduler configuration; no learning-rate or stopping rule was modified after inspection of the locked final-test results.

The final training epoch counts were 27 for FT-Transformer, 49 for LSTM, and 51 for SimpleRNN. These values reflect the frozen development-stage training procedure rather than post-test optimisation. The subsequent checkpoint replay audit reproduced the archived predictions exactly, with a maximum prediction discrepancy of 0.0 kg ha^{-1}. Figure 19 summarises the learning and validation trajectories retained from the completed development runs. The trajectories show both optimisation convergence and substantial fold-dependent variation, particularly under rolling-origin evaluation.

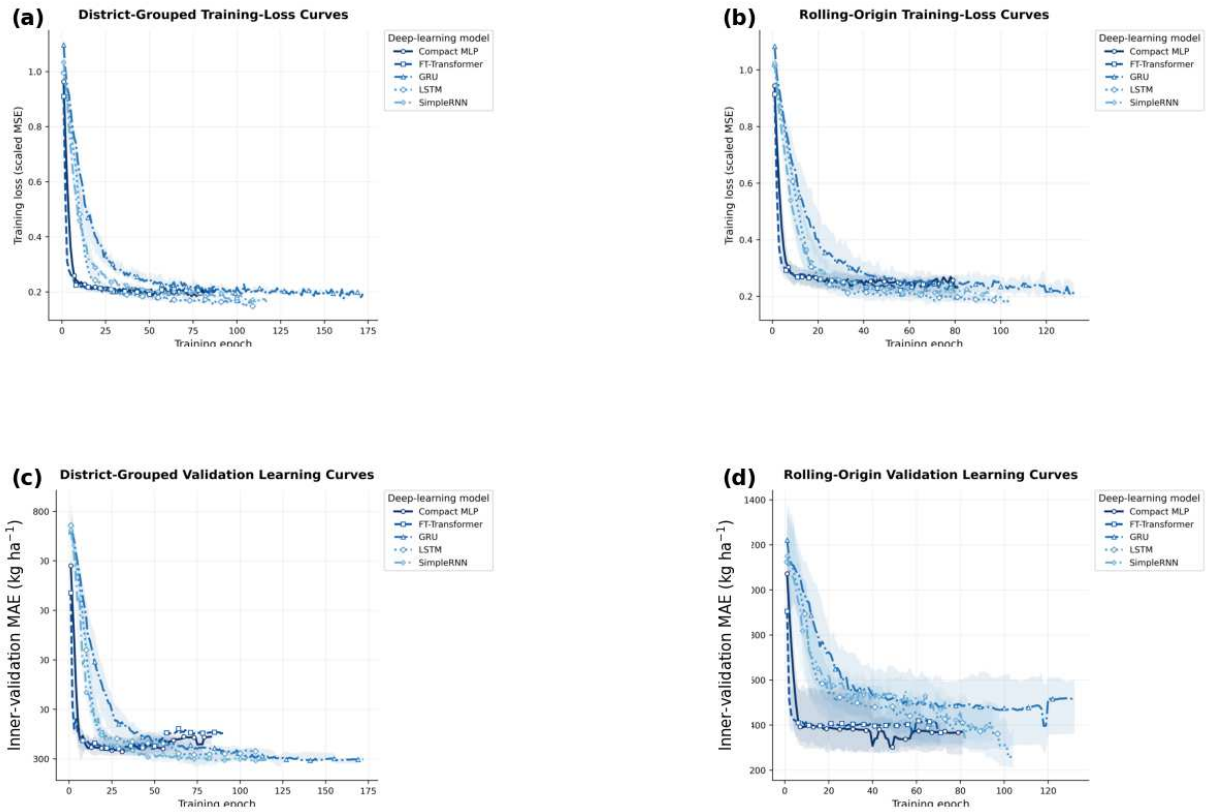


Figure 19. Learning-curve and neural training-history diagnostics under district-grouped and rolling-origin validation.

4.9.2 SimpleRNN versus LSTM Diagnostics

The training diagnostics also provide context for the differing SimpleRNN and LSTM rankings across validation protocols. The canonical LSTM contained 2,505 trainable parameters, whereas SimpleRNN contained 16,929 parameters. Both models received the same seven-month meteorological sequence and eight static historical predictors, but the LSTM used a single gated recurrent layer with hidden size 16, whereas SimpleRNN used two recurrent layers with hidden size 64.

Under district-grouped validation, SimpleRNN achieved an MAE of $304.208\ \mathrm{kgha}^{-1}$, compared with 307.407 kgha^{-1} for LSTM. The difference was only $3.199\ \mathrm{kgha}^{-1}$, or approximately 1.04%, and was not statistically resolved after multiplicity adjustment. This spatial result therefore does not establish a general performance advantage for SimpleRNN.

The ranking reversed under temporal evaluation. LSTM achieved a rolling-origin MAE of $430.873\ \mathrm{kgha}^{-1}$, compared with $459.834\ \mathrm{kgha}^{-1}$ for SimpleRNN, giving LSTM a $28.961\ \mathrm{kgha}^{-1}$ advantage. On the locked final test, LSTM again recorded the lower MAE, at 400.087 versus $406.841\ \mathrm{kgha}^{-1}$, a difference of 6.754 kg ha^{-1}. The relative performance of the two recurrent architectures was therefore protocol dependent rather than uniformly favourable to either model.

Three fixed random seeds—20260721, 20260722, and 20260723—were retained for eligible deep-learning runs, and seed-level predictions were combined using the predefined ensemble procedure. Fold- and seed-level histories were retained rather than selecting a favourable individual run. Best-checkpoint restoration and subsequent replay checks further confirmed that the reported recurrent predictions arose from the archived fitted models rather than from post hoc retraining.

The observed pattern does not provide evidence that LSTM overfitting was definitively responsible for the district-grouped ranking. Instead, the results are consistent with the possibility that the relatively short seven-step sequence and moderate development sample did not consistently reward additional recurrent gating complexity. Conversely, the stronger temporal performance of LSTM suggests that its gated representation may have provided greater stability in some future-year settings. These interpretations remain model- and dataset-specific and do not establish an intrinsic superiority of either recurrent architecture.

Computational parameter counts, checkpoint sizes, archived training times, and storage requirements are reported separately in Section 4.11 to avoid duplicating the computational-efficiency analysis.

4.10 Prediction-Interval Reliability and Calibration

4.10.1 Empirical versus Nominal Coverage

Prediction-interval reliability was evaluated for the five canonical models at nominal coverage levels of 50%, 80%, 90%, and 95%. The analysis used development-stage out-of-fold predictions only. District-grouped calibration excluded observations from the target district and used 592 residuals, whereas rolling-origin calibration excluded observations from the target crop year and used 304 residuals.

Empirical coverage remained close to the requested levels under both validation protocols. Under district-grouped validation, empirical coverage ranged from 50.000% to 50.493% at the 50% nominal level, 80.099% to 80.263% at 80%, 89.967% to 90.296% at 90%, and 95.066% to 95.395% at 95%.

Under rolling-origin validation, empirical coverage ranged from 49.415% to 50.877% at the 50% nominal level, 79.240% to 80.702% at 80%, 89.766% to 90.643% at 90%, and 94.737% to 95.614% at 95%. Figure 20 compares nominal and empirical coverage across the canonical models and both development protocols.

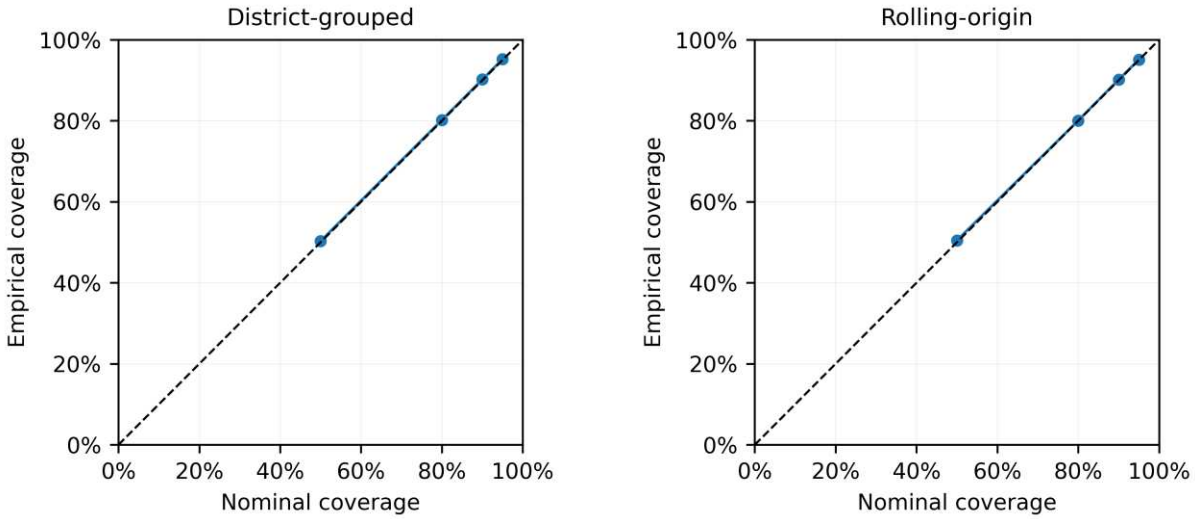


Figure 20. Nominal versus empirical prediction-interval coverage under district-grouped and rolling-origin validation.

4.10.2 Coverage Deviation

Across all models, protocols, and nominal levels, the maximum absolute empirical-minus-nominal coverage deviation was 0.877 percentage points. Under district-grouped validation, signed deviations ranged from 0.000 to +0.493 percentage points at 50% coverage, +0.099 to +0.263 at 80%, -0.033 to +0.296 at 90%, and +0.066 to +0.395 at 95%.

Rolling-origin deviations were slightly larger, ranging from -0.585 to +0.877 percentage points at 50%, -0.760 to +0.702 at 80%, -0.234 to +0.643 at 90%, and -0.263 to +0.614 at 95%. Thus, pooled empirical coverage remained close to nominal coverage despite the more demanding temporal evaluation.

4.10.3 Interval Width and Interval Score

Calibration accuracy was considered jointly with interval sharpness. At 80% nominal coverage, LSTM produced the narrowest district-grouped mean interval width at $971.362\ \text{kgha}^{-1}$, whereas FT-Transformer produced the narrowest rolling-origin width at 1,159.729 kgha^{-1}. At 95% coverage, SimpleRNN produced the narrowest district-grouped interval width at $1{,}719.158\ \text{kgha}^{-1}$, while FT-Transformer again produced the narrowest rolling-origin width at $2{,}254{,}467\ \text{kgha}^{-1}$.

For interval score, SimpleRNN produced the lowest district-grouped values at both 80% and 95% coverage, with scores of $1{,}505.242$ and $2{,}244.764\ \text{kgha}^{-1}$, respectively. Under rolling-origin validation, FT-Transformer produced the lowest interval scores at both levels, at 1,846.822 and 2,675.655 kgha^{-1}. The complete pooled reliability results are reported in Table 19.

Table 19. Cluster-excluded prediction-interval reliability

Protocol	**Model**	**50% coverage**	**80% coverage**	**90% coverage**	**95% coverage**	**80% mean width**	**80% interval score**	**95% mean width**	**95% interval score**
District grouped	Prior two-year moving average	50.000	80.263	90.296	95.066	1,032.372	1,664.718	1,972.592	2,557.199
District grouped	Elastic Net	50.493	80.263	90.132	95.066	999.067	1,618.675	1,874.403	2,428.906

District grouped	FT-Transformer	50.329	80.099	90.296	95.395	1,019.403	1,628.337	1,880.697	2,398.728
District grouped	LSTM	50.493	80.263	90.296	95.395	971.362	1,509.036	1,745.214	2,247.728
District grouped	SimpleRNN	50.164	80.099	89.967	95.066	996.424	1,505.242	1,719.158	2,244.764
Rolling origin	Prior two-year moving average	49.415	80.409	90.351	95.029	1,217.039	1,907.006	2,348.472	2,866.236
Rolling origin	Elastic Net	50.877	80.409	90.058	95.322	1,260.644	1,917.740	2,425.393	2,888.242
Rolling origin	FT-Transformer	50.877	80.702	90.058	94.737	1,159.729	1,846.822	2,254.467	2,675.655
Rolling origin	LSTM	50.877	79.532	89.766	94.737	1,343.801	2,110.260	2,453.935	3,036.708
Rolling origin	SimpleRNN	50.292	79.240	90.643	95.614	1,461.256	2,136.598	2,477.646	2,917.995

Empirical coverage is expressed as a percentage; interval widths are expressed in kg ha^{-1}.

4.10.4 Conditional Reliability

Conditional reliability was examined by stable district, crop year, observed-yield range, and outlier status to assess whether pooled calibration in Table 19 concealed subgroup variation. Coverage and interval width were interpreted jointly.

Reliability varied across spatial, temporal, and yield-based subgroups, but these estimates were less stable than the pooled results because of smaller subgroup sizes. In particular, each high-yield-outlier group contained only five observations and was therefore interpreted descriptively. Accordingly, Table 19 provides the primary evidence of interval calibration, while the conditional analyses serve as supporting sensitivity diagnostics.

4.10.5 Spatial versus Temporal Uncertainty

Prediction intervals were consistently wider under rolling-origin validation than under district-grouped validation for every canonical model at the reported 80% and 95% levels. Interval scores were likewise higher under rolling-origin validation for every model. This pattern indicates that future-year forecasting required greater uncertainty margins than transfer to held-out districts, consistent with the stronger temporal distribution shift identified in the point-prediction and temporal-drift analyses.

The calibration analysis therefore showed that near-nominal pooled coverage was maintained under both protocols, but temporal generalisation required substantially wider intervals and incurred larger interval scores. This distinction between calibration and sharpness is important because similar empirical coverage does not imply equivalent predictive uncertainty across validation settings.

4.11 Computational Efficiency

The canonical neural models contained between 2,505 and 16,929 trainable parameters. LSTM was the smallest neural architecture, with 2,505 parameters, followed by FT-Transformer with 9,185 parameters and SimpleRNN with 16,929 parameters.

Model-storage requirements were also modest for the canonical set. LSTM checkpoints occupied approximately 17-18 KB, FT-Transformer checkpoints approximately 46 KB, and SimpleRNN checkpoints approximately 76 KB. The prior two-year moving-average baseline and Elastic Net required approximately 1.0 and 2.2 KB, respectively.

The recurrent models used seven monthly time steps with 13 time-varying meteorological channels and eight static historical predictors. Elastic Net and FT-Transformer used eight historical predictors, whereas the prior two-year moving-average baseline used a single derived historical summary.

The available computational records are summarised in Table 20. Archived training times originated from different execution stages and are therefore reported descriptively rather than as a controlled hardware-normalised speed benchmark. Deep-model prediction times were retained at the individual-run level but were not aggregated into a cross-model inference-time comparison.

Table 20. Computational and storage characteristics of the canonical models.

Model	Retained input structure	Trainable parameters	Total parameters	Archived training time (s)	Approximate serialized size
Prior two-year moving average	One historical summary	Not applicable	Not applicable	0.000	1.0 KB
Elastic Net	Eight historical predictors	Coefficient-based	Coefficient-based	0.557	2.2 KB
FT-Transformer	Eight historical predictors	9,185	9,185	14.245	46 KB
LSTM	Seven monthly steps × 13 channels plus eight static predictors	2,505	2,505	17.642	17–18 KB
SimpleRNN	Seven monthly steps × 13 channels plus eight static predictors	16,929	16,929	14.494	76 KB

Among the canonical neural models, LSTM achieved the smallest parameter count and checkpoint size, although its archived training time was longer than those of FT-Transformer

and SimpleRNN. These timing differences should not be interpreted as controlled computational-speed comparisons because the retained timing records were obtained at different execution stages.

4.12 Sentinel-2 Seasonal Results

4.12.1 Satellite-Panel Completeness

The Sentinel-2 workflow produced complete early-, peak-, and late-stage district summaries for all 38 stable districts across six crop years from 2017–2018 to 2022–2023. The final satellite panel therefore contained 228 district–crop-year observations.

For every retained stage–index combination, the output registry contained valid-pixel count, mean, median, standard deviation, minimum, maximum, 25th percentile, 75th percentile, interquartile range, and coefficient of variation. No district–crop-year record lacked its corresponding official cultivated area, production, or wheat-yield value.

The complete map inventory contained 30 validated products: early-season NDVI, peak-growth NDVI, late-season NDVI, early-to-peak NDVI change, and peak-to-late NDVI change for each of the six crop years.

4.12.2 Statewide Seasonal Vegetation Patterns

At the statewide mean level, all four spectral indices increased from the early stage to peak growth and subsequently declined during the late stage.

Mean NDVI increased from 0.401 during the early stage to 0.549 at peak growth and declined to 0.241 during the late stage. The corresponding early-to-peak and peak-to-late changes were +0.148 and −0.308, respectively. Mean EVI increased from 0.241 to 0.373 and subsequently declined to 0.152. Mean SAVI increased from 0.232 to 0.348 and declined to 0.156, while mean NDMI increased from 0.018 to 0.164 before declining to −0.085. The statewide seasonal means are presented in Panel A of Table 21.

4.12.3 Annual NDVI Dynamics

Annual NDVI patterns varied across the six Sentinel-2 crop years. The highest statewide peak-growth NDVI was 0.595 in 2019-2020, whereas the lowest was 0.491 in 2018-2019. The largest early-to-peak increase also occurred in 2019-2020, with a statewide mean change of +0.212.

The number of districts recording a positive early-to-peak NDVI change ranged from 23 of 38 in 2018-2019 to 37 of 38 in 2021-2022. All six annual statewide means showed a negative peak-to-late NDVI change. Mean annual wheat yield ranged from $3{,}157.4\ \mathrm{kgha}^{-1}$ in 2018-

2019 to 3,557.1 kg ha^{-1} in 2020-2021. Annual wheat-yield and NDVI summaries are reported in Panel B of Table 21.

Table 21. Sentinel-2 seasonal and annual results

Panel A. Statewide seasonal index means

Index	Early-stage mean	Peak-stage mean	Late-stage mean	Early-to-peak change	Peak-to-late change
NDVI	0.401	0.549	0.241	+0.148	-0.308
EVI	0.241	0.373	0.152	+0.132	-0.221
SAVI	0.232	0.348	0.156	+0.116	-0.192
NDMI	0.018	0.164	-0.085	+0.146	-0.249

Panel B. Annual wheat yield and NDVI dynamics

Crop year	Mean yield (kg ha^{-1})	Early NDVI	Peak NDVI	Late NDVI	Early-to-peak change	Peak-to-late change	Districts with early-to-peak increase
2017–2018	3,348.6	0.397	0.514	0.218	+0.117	−0.296	32/38
2018–2019	3,157.4	0.428	0.491	0.232	+0.063	−0.259	23/38
2019–2020	3,546.1	0.383	0.595	0.261	+0.212	−0.334	34/38
2020–2021	3,557.1	0.391	0.539	0.225	+0.148	−0.314	32/38
2021–2022	3,454.1	0.398	0.580	0.259	+0.182	−0.320	37/38
2022–2023	3,514.9	0.407	0.574	0.252	+0.167	−0.323	34/38

For 2022–2023, the equal-weight district mean peak-growth NDVI was 0.574. Figure 21 presents the corresponding 500-m general-cropland-masked raster distribution and shows substantial spatial heterogeneity in peak-stage vegetation conditions across Madhya Pradesh.

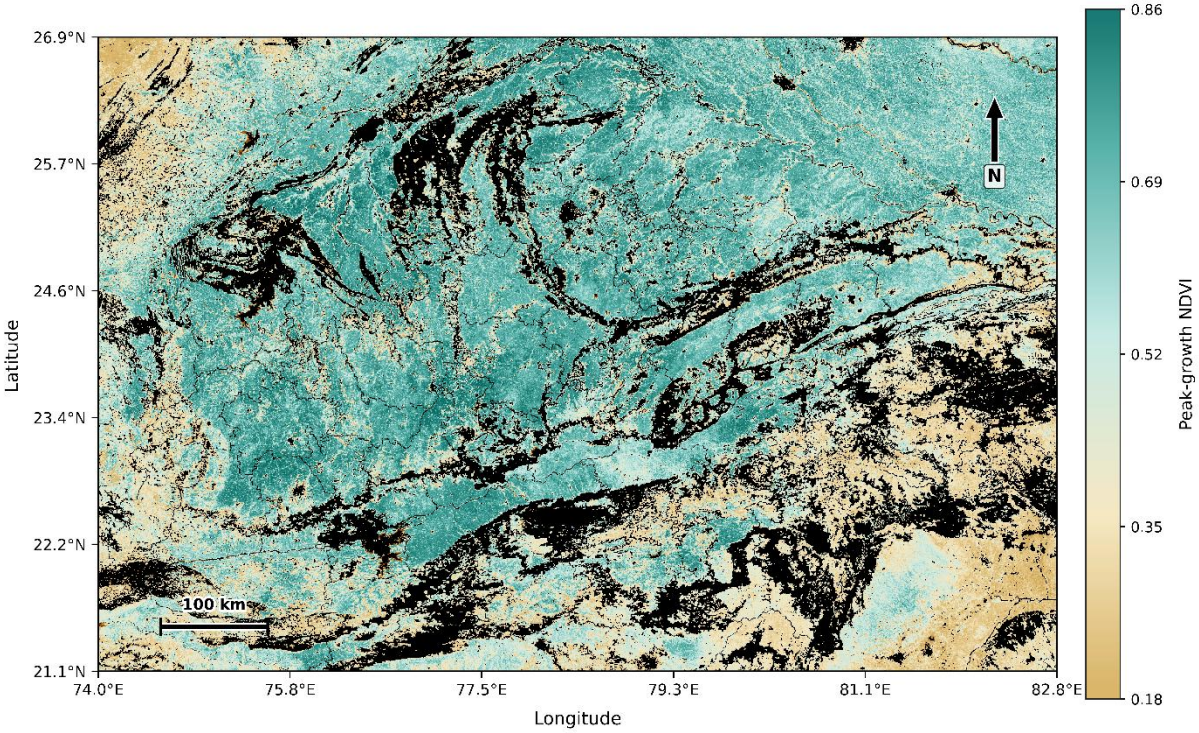


Figure 21. Spatial distribution of peak-growth NDVI over mapped general cropland during 2022–2023.

For the same crop year, 34 of the 38 stable districts recorded a positive early-to-peak NDVI change. The district-summary mean increased from 0.407 during the early stage to 0.574 at peak growth, corresponding to a mean change of +0.167. Figure 22 shows the spatial distribution of this early-to-peak increase.

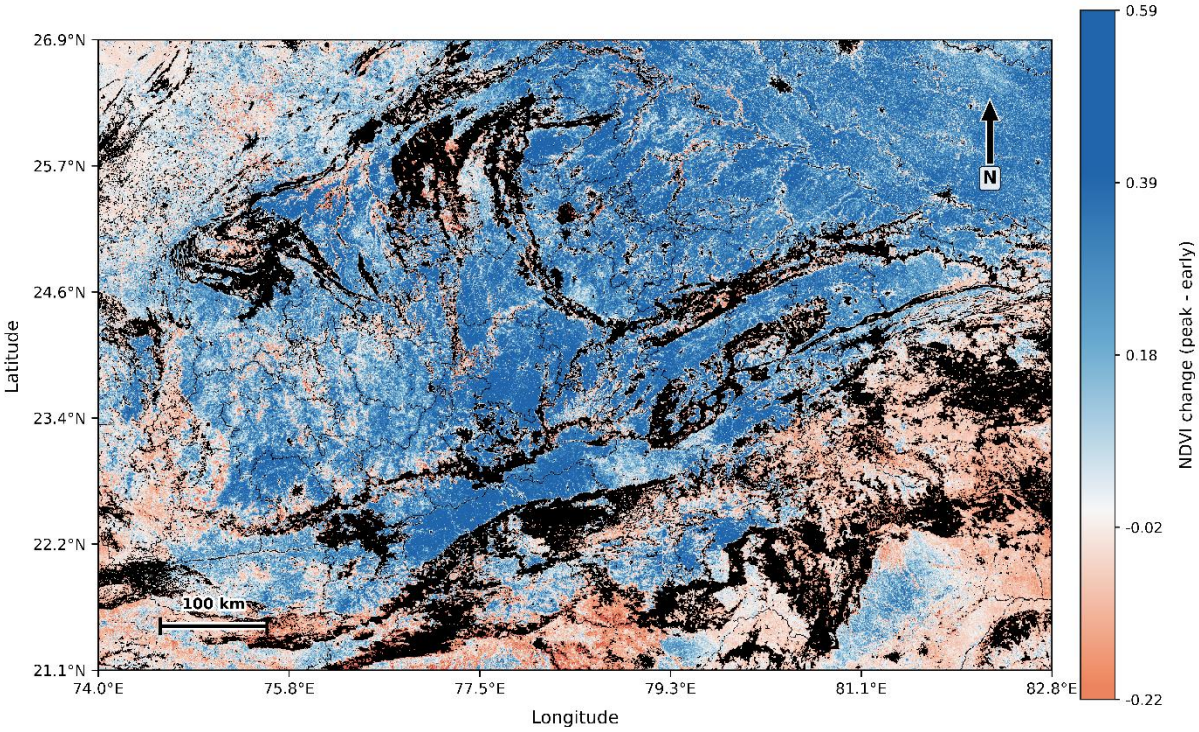


Figure 22. Spatial distribution of early-to-peak NDVI change over mapped general cropland during 2022–2023.

Peak-to-late NDVI change was negative at the statewide district-summary level, decreasing from 0.574 at peak growth to 0.252 during the late stage, corresponding to a mean change of −0.323. Figure 23 shows the spatial distribution of the late-season decline.

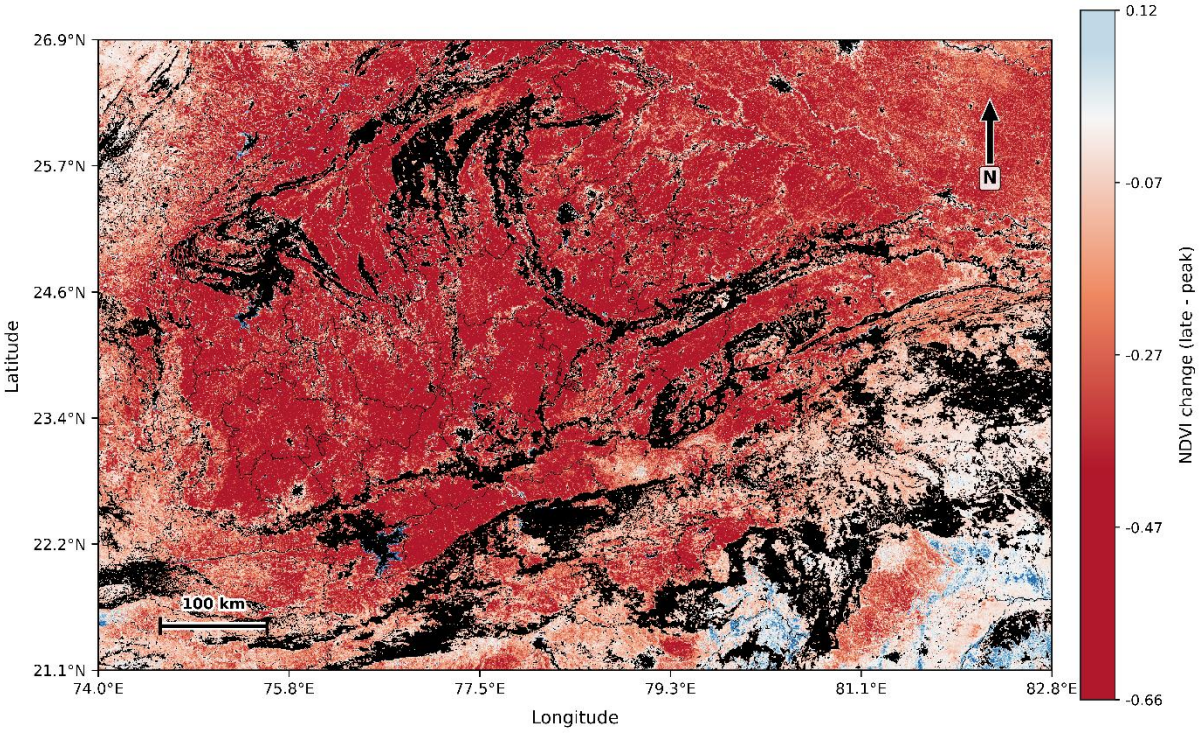


Figure 23. Spatial distribution of peak-to-late NDVI change over mapped general cropland during 2022–2023.

4.12.4 Sentinel-2-Yield Associations

Across the 228 district-crop-year observations, late-stage NDMI showed the strongest pooled association with official wheat yield, with Pearson $r = 0.442$ and Spearman $\rho = 0.507$. Peak-stage NDMI showed a similar positive association ($r = 0.439, \rho = 0.442$), while early-stage NDMI was also positively associated with yield ($r = 0.335, \rho = 0.426$).

Among the peak-stage vegetation indices, peak NDVI showed Pearson and Spearman correlations of 0.375 and 0.376, respectively, followed by peak SAVI ($r = 0.347, \rho = 0.358$) and peak EVI ($r = 0.329, \rho = 0.335$). In contrast, peak-to-late changes were negatively associated with yield for NDVI ($r = -0.367, \rho = -0.333$), SAVI ($r = -0.352, \rho = -0.328$), NDMI ($r = -0.333, \rho = -0.315$), and EVI ($r = -0.333$, $\rho = -0.309$).

These relationships are interpreted as descriptive district-level associations only. The satellite analysis used a general-cropland mask rather than a wheat-specific classification, and the Sentinel-2 variables were not incorporated into the frozen canonical forecasting models. The associations therefore do not establish causality, wheat-classification accuracy, or an improvement in predictive performance attributable to Sentinel-2.

5. Discussion

The results show that district-level Rabi wheat-yield prediction in Madhya Pradesh depended more strongly on historical productivity and validation setting than on model complexity. Under district-grouped validation, SimpleRNN achieved the lowest MAE of 304.208 kg ha^{-1}, but its improvement over the prior two-year moving-average baseline was only 2.78% and was not statistically resolved after multiplicity correction. Under rolling-origin validation, FT-Transformer ranked first with an MAE of 367.610 kg ha^{-1}, only 0.82% better than the persistence baseline. On the protected final test, Elastic Net achieved the lowest MAE of 368.498 kg ha^{-1}, representing a 3.79% reduction relative to the moving-average baseline. This result is consistent with evidence that deep learning does not necessarily outperform well-regularised methods on moderate-sized structured tabular datasets [74]. The changing rankings across protocols further demonstrate that geographical transfer and future-year forecasting represent different generalisation problems.

The protected-test advantage of Elastic Net is therefore an important result rather than evidence of model inadequacy. The complete archive contained 950 district–crop-year observations, but

repeated measurements from 38 districts meant that these were not 950 independent samples, while only 608 observations formed the common development panel. Under these conditions, coefficient shrinkage can provide lower-variance estimates than highly flexible representations when recent productivity already explains a substantial proportion of yield variation. This interpretation is also consistent with agricultural studies showing that interpretable models can reveal meaningful yield structure without requiring unnecessary complexity [75]. Consequently, deep architectures should be judged against strong persistence and regularised baselines rather than assumed to provide automatic gains.

Historical productivity was the dominant predictive information source. Adding eight historical-yield predictors to the matched LSTM reduced MAE by 50.1% under district-grouped validation and 59.4% under rolling-origin validation, while neutralising the historical block produced the largest performance deterioration across the fitted models. These variables should not, however, be interpreted causally. Previous yields may indirectly encode persistent differences in irrigation infrastructure, soil quality, management intensity, technology and local production systems. Irrigation, for example, substantially modifies wheat productivity and heat sensitivity in India [76], while extreme heat can produce important yield risks that are not completely represented by historical persistence [77]. Crop-management timing can similarly influence resilience and productivity [78]. Thus, historical yield functions as a strong composite predictive indicator rather than a direct measurement of the agronomic mechanisms responsible for productivity.

Weather nevertheless contributed useful information. Sequence-length perturbation showed that reducing the effective October–April sequence generally increased recurrent-model error, particularly when only one to four months were retained. Integrated Gradients and frozen-model perturbation also identified relative humidity, heat stress, wind and particular late-season months as influential, although ranking agreement weakened under rolling-origin validation. This protocol dependence supports cautious interpretation of neural explanations. Previous interpretable machine-learning research in Indian wheat systems likewise demonstrates that environmental and management drivers can interact strongly in determining yield variability [79]. Heat effects can additionally depend on sowing date and developmental timing [80], which could not be evaluated directly here because consistent sowing-date observations were unavailable.

The SimpleRNN–LSTM comparison illustrates why model rankings should not be interpreted from a single validation protocol. SimpleRNN was only 3.199 kg ha^{-1} better than LSTM under district-grouped validation, approximately 1.04%, and the difference was not statistically

resolved. The ranking reversed under rolling-origin validation, where LSTM was 28.961 kg ha^{-1} better, and LSTM also slightly outperformed SimpleRNN on the protected test. The seven-month sequence may therefore not consistently reward additional gating complexity, but the evidence does not establish either architecture as intrinsically superior. More importantly, agricultural responses can depend on compound weather conditions that differ between seasons; hot, dry and windy events have been associated with substantial winter-wheat losses elsewhere [81], while exceptional climatic combinations can produce abrupt yield failures that standard models struggle to represent [82].

The failure analyses reinforce this interpretation. All models were substantially less accurate for the small number of high-yield observations, indicating regression towards the dominant central yield range. Rolling-origin validation also produced stronger negative bias and greater temporal instability than district-grouped validation. Comparatively wet conditions were associated with increased errors under temporal generalisation, whereas the highest heat-stress and longest dry-spell strata were not systematically the most difficult. These findings should therefore not be reduced to single-variable causal explanations. Global crop-modelling evidence similarly indicates that climate-related agricultural impacts may emerge differently across environments and modelling systems [83], while longer-term genetic improvement may itself be challenged by changing climatic conditions [84].

Prediction uncertainty showed a similar spatial–temporal distinction. Empirical coverage remained close to nominal levels, with a maximum absolute deviation of only 0.877 percentage points, as illustrated by the calibration curves in Figure 20. However, rolling-origin intervals were consistently wider and produced higher interval scores, demonstrating that future-year forecasting required larger uncertainty margins than transfer to held-out districts. Climate variables can provide substantial forecasting information late in the wheat season [85], but concurrent extremes across major wheat-growing regions demonstrate why uncertainty should remain explicit in operational forecasting [86]. The present residual-calibration approach provided a common comparison across all model families; MC Dropout, deep ensembles, quantile regression and alternative conformal approaches remain appropriate targets for future comparative evaluation rather than retrospective additions to the frozen benchmark.

The study also strengthens interpretability through SHAP interactions, bootstrap stability and cross-method comparison. Elastic Net historical-feature rankings remained unchanged across 1,000 district-cluster bootstrap resamples, whereas recurrent Integrated-Gradients rankings were less stable under temporal validation. This distinction is agronomically plausible because the importance of climatic conditions may vary across developmental stages [87]. Heat effects

can differ before and after flowering [88], while drought responses also depend on phenological stage [89]. Accordingly, SHAP and Integrated Gradients should be interpreted as descriptions of fitted-model dependence, not evidence of causal physiological effects.

The Sentinel-2 results provide complementary recent-period evidence rather than an alternative forecasting model. Late-stage NDMI showed the strongest satellite–yield association, but the analysis covered only six seasons and used a general-cropland rather than wheat-specific mask. Figures 21–23 therefore describe spatial crop-condition patterns rather than wheat-specific yield predictions. Longer-term wheat improvement under optimum, drought and heat environments further illustrates that observed productivity reflects interacting genetic, environmental and management influences [90].

Overall, the framework is most useful as a leakage-aware and reproducible agricultural forecasting benchmark rather than as evidence that one architecture is universally optimal. Its principal limitations are the moderate effective sample size, single-state and single-crop scope, unavailable soil, irrigation, fertiliser, groundwater and sowing-date variables, one meteorological reanalysis product, six Sentinel-2 seasons, and absence of external geographical validation. FT-Transformer represented one compact tabular attention model; TabNet, SAINT, NODE, TabTransformer, TabPFN, Temporal Fusion Transformer, Informer, Autoformer, PatchTST, TimesNet, N-HiTS and Chronos were not benchmarked and should form part of future prospective comparisons. External validation across Punjab, Haryana, Maharashtra and other wheat-producing regions, followed by extension to rice, soybean and maize with crop-specific seasonal definitions, is required before broader generalisation. Future work should additionally incorporate soil and irrigation information, wheat-specific remote-sensing masks, genuinely prospective unseen years, strict cold-start districts, joint spatiotemporal holdouts, alternative uncertainty methods, and models trained explicitly for different in-season forecasting dates.

Conclusion

This study developed an explainable and leakage-aware framework for district-level Rabi wheat-yield prediction in Madhya Pradesh using harmonised agricultural records, stable historical district units, ERA5-Land meteorology, historical-yield predictors, and complementary Sentinel-2 indicators. Model rankings varied across validation settings, confirming that geographical transfer and future-year forecasting are distinct prediction problems. SimpleRNN achieved the lowest district-grouped MAE of 304.208 kg ha^{-1}, FT-Transformer achieved the lowest rolling-origin MAE of 367.610 kg ha^{-1}, and Elastic Net

performed best on the protected final test with an MAE of 368.498 kg ha^{-1}, improving upon the prior two-year moving-average baseline by 3.79%. Explainability, matched ablation, and frozen-model perturbation consistently showed that recent district-level productivity history was the dominant predictive information source; adding historical-yield features reduced LSTM MAE by 50.1% under district-grouped validation and 59.4% under rolling-origin validation. Weather remained informative, but its contribution was less stable under temporal generalisation. Prediction intervals remained close to nominal coverage, with a maximum absolute deviation of 0.877 percentage points, but were wider under rolling-origin validation, indicating greater uncertainty in future-year forecasting. High-yield observations were consistently more difficult to predict, while Sentinel-2 analysis identified late-stage NDMI as the strongest satellite–yield correlate, although these associations remained descriptive. Overall, the findings show that strong historical baselines, regularised models, leakage-aware validation, protected testing, interpretability, uncertainty assessment, and reproducibility controls are more important than model complexity alone. Future work should include additional wheat-producing regions, wheat-specific remote-sensing masks, soil and irrigation variables, strict cold-start and joint spatiotemporal validation, modern tabular and temporal architectures, and prospective testing on genuinely unseen years and regions.

Author Contributions: Viswavardhan Reddy Karna: conceived and designed the study, developed the methodology, performed the data curation and analysis, conducted the machine-learning and deep-learning experiments, performed the explainability, validation, robustness and uncertainty analyses, and wrote the original manuscript. Neethu S: contributed to the research design, data analysis, interpretation of results, and manuscript review. Supreeth S: contributed to data processing, computational analysis, and manuscript review. Sarala D V: contributed to data curation, analysis, interpretation of results, and manuscript review. Sunitha T: contributed to data preparation, analysis, and manuscript review. Vishnu Vardhana Reddy Karna: contributed to the research methodology, interpretation of results, and manuscript review. All authors reviewed and approved the final manuscript.

Funding Declaration: This work received no external funding.

Data Availability: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of interest: The authors declare no competing interests.

References

[1] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* 177 (Oct. 2020), Article 105709.

[2] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin, Kyo-Moon Shim, James S. Gerber, Vangimalla R. Reddy, and Soo-Hyung Kim. 2016. Random forests for global and regional crop yield predictions. *PLOS ONE* 11, 6 (June 2016), Article e0156571.

[3] Saeed Khaki and Lizhi Wang. 2019. Crop yield prediction using deep neural networks. *Frontiers in Plant Science* 10 (May 2019), Article 621.

[4] Jiaxuan You, Xiaocheng Li, Melvin Low, David B. Lobell, and Stefano Ermon. 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI '17)*. AAAI Press, 4559–4565.

[5] Andrew Crane-Droesch. 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters* 13, 11 (Nov. 2018), Article 114003.

[6] Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V. Archontoulis. 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports* 11 (Jan. 2021), Article 1606.

[7] David B. Lobell and Senthold Asseng. 2017. Comparing estimates of climate change impacts from process-based and statistical crop models. *Environmental Research Letters* 12, 1 (Jan. 2017), Article 015001.

[8] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 8 (Aug. 2017), 913–929.

[9] Christoph Bergmeir and José M. Benítez. 2012. On the use of cross-validation for time-series predictor evaluation. *Information Sciences* 191 (May 2012), 192–213.

[10] Vítor Cerqueira, Luís Torgo, and Igor Mozetič. 2020. Evaluating time-series forecasting models: An empirical study on performance-estimation methods. *Machine Learning* 109, 11 (Nov. 2020), 1997–2028.

[11] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N. Athanasiadis. 2021. Machine learning for large-scale crop yield forecasting. *Agricultural Systems* 187 (2021), Article 103016.

[12] Michele Meroni, François Waldner, Lorenzo Seguini, Hervé Kerdiles, and Felix Rembold. 2021. Yield forecasting with machine learning and small data: What gains for grains? *Agricultural and Forest Meteorology* 308–309 (2021), Article 108555.

[13] Yaping Cai, Kaiyu Guan, David B. Lobell, Alex B. Potgieter, Shaowen Wang, Jian Peng, Tao Xu, Senthold Asseng, Yao Zhang, Liangzhi You, and Bin Peng. 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* 274 (2019), 144–159.

[14] Diego Gómez, Pablo Salvador, Julia Sanz, and José Luis Casanova. 2021. Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico. *Agricultural and Forest Meteorology* 300 (2021), Article 108317.

[15] Penghao Feng, Bin Wang, De Li Liu, Cathryn Waters, Dong Xiao, Lingling Shi, and Qiang Yu. 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agricultural and Forest Meteorology* 285–286 (2020), Article 107922.

[16] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N. Athanasiadis. 2022. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research* 276 (2022), Article 108377.

[17] Djavan De Clercq and Adam Mahdi. 2024. Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. *Agricultural Systems* 220 (2024), Article 104099.

[18] Donghoon Lee, Frank Davenport, Shraddhanand Shukla, Greg Husak, Chris Funk, and James Verdin. 2024. Contrasting performance of panel and time-series data models for subnational crop forecasting in Sub-Saharan Africa. *Agricultural and Forest Meteorology* 359 (2024), Article 110213.

[19] Merryn L. Hunt, George Alan Blackburn, Luis Carrasco, John W. Redhead, and Clare S. Rowland. 2019. High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment* 233 (2019), Article 111410.

[20] Joan Segarra, José Luis Araus, and Shawn C. Kefauver. 2022. Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield. *International Journal of Applied Earth Observation and Geoinformation* 107 (2022), Article 102697.

[21] Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. 2019. 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* 10, 2 (2019), 225–232.

[22] Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten F. Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* 11 (2020), Article 4540.

[23] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[24] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297.

[25] Jakob Gawlikowski, Cedrique R. N. Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 56, Supplement 1 (2023), 1513–1589.

[26] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 2 (April 2005), 301–320.

[27] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (April 2006), 3–42.

[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780.

[29] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. Curran Associates, 18932–18943.

[30] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Curran Associates, 4765–4774.

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML '17)*. Proceedings of Machine Learning Research, Vol. 70, 3319–3328.

[32] Alfredo R. Huete. 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* 25, 3 (Aug. 1988), 295–309.

[33] Alfredo Huete, Kamel Didan, Tomoaki Miura, E. Patricia Rodriguez, Xiang Gao, and Laerte G. Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* 83, 1–2 (Nov. 2002), 195–213.

[34] Bo-Cai Gao. 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment* 58, 3 (Dec. 1996), 257–266.

[35] Huiren Tian, Pengxin Wang, Kevin Tansey, Jingqi Zhang, Shuyu Zhang, and Hongmei Li. 2021. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. *Agricultural and Forest Meteorology* 310 (2021), Article 108629.

[36] Raí A. Schwalbert, Telmo J. C. Amado, Geomar M. Corassa, Luan P. Pott, P. V. V. Prasad, and Ignacio A. Ciampitti. 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology* 284 (2020), Article 107886.

[37] Keyhan Gavahi, Peyman Abbaszadeh, and Hamid Moradkhani. 2021. DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications* 184 (2021), Article 115511.

[38] Peyman Abbaszadeh, Keyhan Gavahi, Atieh Alipour, Proloy Deb, and Hamid Moradkhani. 2022. Bayesian multi-modeling of deep neural nets for probabilistic crop yield prediction. *Agricultural and Forest Meteorology* 314 (2022), Article 108773.

[39] Florian Huber, Artem Yushchenko, Benedikt Stratmann, and Volker Steinhage. 2022. Extreme Gradient Boosting for yield estimation compared with deep learning approaches. *Computers and Electronics in Agriculture* 202 (2022), Article 107346.

[40] Dilli Paudel, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, and Ioannis N. Athanasiadis. 2023. Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture* 206 (2023), Article 107663.

[41] Johann Desloires, Dino Ienco, and Antoine Botrel. 2023. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Computers and Electronics in Agriculture* 209 (2023), Article 107807.
[42] Yilin Zhu, Sensen Wu, Mengjiao Qin, Zhiyi Fu, Yi Gao, Yuanyuan Wang, and Zhenhong Du. 2022. A deep learning crop model for adaptive yield estimation in large areas. *International Journal of Applied Earth Observation and Geoinformation* 110 (2022), Article 102828.
[43] Mengjia Qiao, Xiaohui He, Xijie Cheng, Panle Li, Haotian Luo, Lehan Zhang, and Zhihui Tian. 2021. Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks. *International Journal of Applied Earth Observation and Geoinformation* 102 (2021), Article 102436.
[44] V. Sagan, M. Maimaitijiang, S. Bhadra, M. Maimaitiyiming, D. R. Brown, P. Sidike, and F. B. Fritschi. 2021. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* 174 (2021), 265–281.
[45] Huiren Tian, Pengxin Wang, Kevin Tansey, Dong Han, Jingqi Zhang, Shuyu Zhang, and Hongmei Li. 2021. A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China. *International Journal of Applied Earth Observation and Geoinformation* 102 (2021), Article 102375.
[46] Yahui Guo, Yi Xiao, Fanghua Hao, Xuan Zhang, Jiahao Chen, Kirsten de Beurs, Yuhong He, and Yongshuo H. Fu. 2023. Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation* 124 (2023), Article 103528.
[47] Dong Han, Pengxin Wang, Kevin Tansey, Yue Zhang, and Hongmei Li. 2024. A graph-based deep learning framework for field scale wheat yield estimation. *International Journal of Applied Earth Observation and Geoinformation* 129 (2024), Article 103834.
[48] Huiren Tian, Pengxin Wang, Kevin Tansey, Jie Wang, Wenting Quan, and Junming Liu. 2024. Attention mechanism-based deep learning approach for wheat yield estimation and uncertainty analysis from remotely sensed variables. *Agricultural and Forest Meteorology* 356 (2024), Article 110183.
[49] Huimin Zhuang, Zhao Zhang, Fei Cheng, Jichong Han, Yuchuan Luo, Liangliang Zhang, Juan Cao, Jing Zhang, Bangke He, Jialu Xu, and Fulu Tao. 2024. Integrating data assimilation, crop model, and machine learning for winter wheat yield forecasting in the North China Plain. *Agricultural and Forest Meteorology* 347 (2024), Article 109909.

[50] Chenxi Lu, Guoyong Leng, Xiaoyong Liao, Haiyang Tu, Jiali Qiu, Ji Li, Shengzhi Huang, and Jian Peng. 2024. In-season maize yield prediction in Northeast China: The phase-dependent benefits of assimilating climate forecast and satellite observations. *Agricultural and Forest Meteorology* 358 (2024), Article 110242.

[51] Weiguo Yu, Gaoxiang Yang, Dong Li, Hengbiao Zheng, Xia Yao, Yan Zhu, Weixing Cao, Lin Qiu, and Tao Cheng. 2023. Improved prediction of rice yield at field and county levels by synergistic use of SAR, optical and meteorological data. *Agricultural and Forest Meteorology* 342 (2023), Article 109729.

[52] Shangrong Wu, Peng Yang, Jianqiang Ren, Zhongxin Chen, and He Li. 2021. Regional winter wheat yield estimation based on the WOFOST model and a novel VW-4DEnSRF assimilation algorithm. *Remote Sensing of Environment* 255 (2021), Article 112276.

[53] Guanyuan Shuai and Bruno Basso. 2022. Subfield maize yield prediction improves when in-season crop water deficit is included in remote sensing imagery-based models. *Remote Sensing of Environment* 272 (2022), Article 112938.

[54] Matteo G. Ziliani, Muhammad U. Altaf, Bruno Aragon, Rasmus Houborg, Trenton E. Franz, Yang Lu, Justin Sheffield, Ibrahim Hoteit, and Matthew F. McCabe. 2022. Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model. *Agricultural and Forest Meteorology* 313 (2022), Article 108736.

[55] Deborah V. Gaso, Allard de Wit, Andres G. Berger, and Lammert Kooistra. 2021. Predicting within-field soybean yield variability by coupling Sentinel-2 leaf area index with a crop growth model. *Agricultural and Forest Meteorology* 308–309 (2021), Article 108553.

[56] Yang Chen, Randall J. Donohue, Tim R. McVicar, François Waldner, Gonzalo Mata, Noboru Ota, Alireza Houshmandfar, Kavina Dayal, and Roger A. Lawes. 2020. Nationwide crop yield estimation based on photosynthesis and meteorological stress indices. *Agricultural and Forest Meteorology* 284 (2020), Article 107872.

[57] Zhenhai Li, James Taylor, Hao Yang, Raffaele Casa, Xiuliang Jin, Zhenhong Li, Xiaoyu Song, and Guijun Yang. 2020. A hierarchical interannual wheat yield and grain protein prediction model using spectral vegetative indices and meteorological data. *Field Crops Research* 248 (2020), Article 107711.

[58] Yuval Sadeh, Xuan Zhu, David Dunkerley, Jeffrey P. Walker, Yang Chen, and Karine Chenu. 2024. Versatile crop yield estimator. *Agronomy for Sustainable Development* 44 (2024), Article 42.

[59] Ahmed M. S. Kheir, Ajit Govind, Vinay Nangia, Maher A. El-Maghraby, Abdelrazek Elnashar, Mukhtar Ahmed, Hesham Aboelsoud, Rania Mostafa, and Til Feike. 2025.

Hybridization of process-based models, remote sensing, and machine learning for enhanced spatial predictions of wheat yield and quality. *Computers and Electronics in Agriculture* 234 (2025), Article 110317.

[60] Muhammad Ashfaq, Imran Khan, Dilawar Shah, Shujaat Ali, and Muhammad Tahir. 2025. Predicting wheat yield using deep learning and multi-source environmental data. *Scientific Reports* 15 (2025), Article 26446.

[61] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2 (2020), 56–67.

[62] Kjersti Aas, Martin Jullum, and Anders Løland. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* 298 (2021), Article 103502.

[63] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (2019), 206–215.

[64] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113, 523 (2018), 1094–1111.

[65] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. Predictive inference with the jackknife+. *The Annals of Statistics* 49, 1 (2021), 486–507.

[66] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102, 477 (March 2007), 359–378.

[67] Directorate of Economics and Statistics, Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India. 2024. District-wise, season-wise crop production statistics from 1997. *Open Government Data Platform India*. https://www.data.gov.in/resource/district-wise-season-wise-crop-production-statistics-1997

[68] Survey of India, Government of India. 2026. Administrative Boundary Database: Entire country up to district level with headquarters. Product Code OVSF/1M/7, scale 1:1,000,000, Shapefile. *Survey of India Online Maps Portal*. https://onlinemaps.surveyofindia.gov.in/Digital_Products.aspx

[69] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J.

Rodríguez-Fernández, E. Zsoter, C. Buontempo, and J.-N. Thépaut. 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data* 13, 9 (Sept. 2021), 4349–4383. https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_DAILY_AGGR

[70] European Union, European Space Agency, and Copernicus Programme. 2017. Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A surface reflectance. Google Earth Engine collection COPERNICUS/S2_SR_HARMONIZED. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED

[71] V. J. Pasquarella, C. F. Brown, W. Czerwinski, and W. J. Rucklidge. 2023. Comprehensive quality assessment of optical satellite imagery using weakly supervised video learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '23)*. IEEE, 2125–2135. https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_CLOUD_SCORE_PLUS_V1_S2_HARMONIZED

[72] C. F. Brown, S. P. Brumby, B. Guzder-Williams, T. Birch, S. Brooks Hyde, J. Mazzariello, W. Czerwinski, V. J. Pasquarella, R. Haertel, S. Ilyushchenko, K. Schwehr, M. Weisse, F. Stolle, C. Hanson, O. Guinan, R. Moore, and A. M. Tait. 2022. Dynamic World, near real-time global 10 m land-use and land-cover mapping. *Scientific Data* 9 (June 2022), Article 251. https://developers.google.com/earth-engine/datasets/catalog/GOOGLE_DYNAMICWORLD_V1

[73] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. 2022. ESA WorldCover 10 m 2021 v200. Zenodo, version v200. https://developers.google.com/earth-engine/datasets/catalog/ESA_WorldCover_v200

[74] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.

[75] Edward J. Jones, Thomas F. A. Bishop, Brendan P. Malone, Patrick J. Hulme, Brett M. Whelan, and Patrick Filippi. 2022. Identifying causes of crop yield variability with interpretive machine learning. *Computers and Electronics in Agriculture* 192 (2022), Article 106632.

[76] Esha Zaveri and David B. Lobell. 2019. The role of irrigation in changing wheat yields and heat sensitivity in India. *Nature Communications* 10 (2019), Article 4144.

[77] Mariam Zachariah, Arpita Mondal, and Amir AghaKouchak. 2021. Probabilistic assessment of extreme heat stress on Indian wheat yields under climate change. *Geophysical Research Letters* 48, 20 (2021), Article e2021GL094702.

[78] Andrew J. McDonald, Balwinder-Singh, Alwin Keil, Amit Srivastava, Peter Craufurd, Avinash Kishore, Virender Kumar, Gokul Paudel, Sudhanshu Singh, A. K. Singh, R. K. Sohane, and R. K. Malik. 2022. Time management governs climate resilience and productivity in the coupled rice–wheat cropping systems of eastern India. *Nature Food* 3 (2022), 542–551.

[79] Hari Sankar Nayak, João Vasco Silva, Chiter Mal Parihar, Timothy J. Krupnik, D. R. Sena, S. K. Kakraliya, H. S. Jat, H. S. Sidhu, P. C. Sharma, M. L. Jat, and T. B. Sapkota. 2022. Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India. *Field Crops Research* 287 (2022), Article 108640.

[80] Raj K. Jat, Parvinder Singh, Mangi L. Jat, Mahendra Dia, Harminder S. Sidhu, Shankar L. Jat, Deepak Bijarniya, Hanuman S. Jat, Chhiter M. Parihar, Uttam Kumar, and Santiago L. Ridaura. 2018. Heat stress and yield stability of wheat genotypes under different sowing dates across agro-ecosystems in India. *Field Crops Research* 218 (2018), 33–50.

[81] Haidong Zhao, Lina Zhang, M. B. Kirkham, Stephen M. Welch, John W. Nielsen-Gammon, Guihua Bai, Jiebo Luo, Daniel A. Andresen, Charles W. Rice, Nenghan Wan, Romulo P. Lollato, Dianfeng Zheng, Prasanna H. Gowda, and Xiaomao Lin. 2022. U.S. winter wheat yield loss attributed to compound hot-dry-windy events. *Nature Communications* 13 (2022), Article 7233.

[82] Tamara Ben-Ari, Julien Boé, Philippe Ciais, Remi Lecerf, Marijn van der Velde, and David Makowski. 2018. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications* 9 (2018), Article 1627.

[83] Jonas Jägermeyr, Christoph Müller, Alex C. Ruane, et al. 2021. Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nature Food* 2 (2021), 873–885.

[84] Tianyi Zhang, Yong He, Ron DePauw, Zhenong Jin, David Garvin, Xu Yue, Weston Anderson, Tao Li, Xin Dong, Tao Zhang, and Xiaoguang Yang. 2022. Climate change may outpace current wheat breeding yield improvements in North America. *Nature Communications* 13 (2022), Article 5591.

[85] Kavina Dayal, Jaclyn N. Brown, François Waldner, Roger Lawes, Zvi Hochman, Randall Donohue, Heidi Horan, and Yang Chen. 2020. Climate drivers provide valuable insights into late season prediction of Australian wheat yield. *Agricultural and Forest Meteorology* 295 (2020), Article 108202.

[86] Andrea Toreti, Ottmar Cronie, and Matteo Zampieri. 2019. Concurrent climate extremes in the key wheat producing regions of the world. *Scientific Reports* 9 (2019), Article 5493.

[87] Florian Schierhorn, Max Hofmann, Taras Gagalyuk, Igor Ostapchuk, and Daniel Müller. 2021. Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. *Climatic Change* 169 (2021), Article 39.

[88] Najeeb Ullah, Brian Collins, John T. Christopher, Troy M. Frederiks, and Karine Chenu. 2024. Pre- and post-flowering impacts of natural heatwaves on yield components in wheat. *Field Crops Research* 316 (2024), Article 109489.

[89] Cevat Eser, Suleyman Soylu, and Hakan Ozkan. 2024. Drought responses of traditional and modern wheats in different phenological stages. *Field Crops Research* 305 (2024), Article 109201.

[90] Suchismita Mondal, Somak Dutta, Leonardo Crespo-Herrera, Julio Huerta-Espino, Hans J. Braun, and Ravi P. Singh. 2020. Fifty years of semi-dwarf spring wheat breeding at CIMMYT: Grain yield progress in optimum, drought and heat stress environments. *Field Crops Research* 250 (2020), Article 107757.