Leakage-safe, explainable multi-source machine learning for district-level rice yield forecasting: a ground-truth simulation benchmark for Northeast Thailand

Thanatchon Chongmankhong

A PREPRINT · November 7, 2026

Abstract

District-level rice yield forecasts support food-security monitoring, disaster relief and crop insurance, yet machine-learning studies combining satellite, weather and economic data rarely quantify evaluation leakage or verify that explanations reflect real mechanisms. We built a process-informed simulation benchmark of main-season rice in 200 districts of Northeast Thailand over 18 seasons, whose known data-generating process encodes drought, flood, heat, disease, salinity, price and competing-crop effects, persistent management and satellite-only pest outbreaks, and emits cloud-contaminated satellite indices, weather, prices and noisy map layers together with the true yield components. Eight learners were compared under a leakage-safe rolling origin with 81 features in eight groups. Multi-source learners reduced the RMSE from 47.1 kg rai^{-1} (climatology) to 27.4 (LightGBM) and 29.9 (the validation-selected LightGBM–Ridge ensemble); Holm-adjusted Diebold–Mariano tests over eight test seasons were not significant, but the gains replicated in ten independent worlds. Random cross-validation understated the error by 10–21% and concealed catastrophic failures of the linear learner, which a post-hoc analysis traced to row-wise leave-one-out tuning of its penalty. Satellite features were the only group whose removal materially degraded LightGBM (+25.0%), and their value emerged only from October. Grouped SHAP values recovered strong mechanisms (drought, $r = 0.84$) but misattributed heat stress to moisture and absorbed the technology trend into prices; a fragile model produced confident but wrong attributions. Rolling conformal 80% intervals covered 78.3% of outcomes. All results concern the synthetic benchmark; code and a real-data pathway are provided.

Keywords rice yield forecasting · remote sensing · machine learning · data leakage · explainable AI · conformal prediction · simulation study · Thailand

1 Introduction

Rice is the backbone of agriculture in Northeast Thailand, where most of the crop is grown as a single, rainfed wet-season (main-season) crop on the sandy and partly saline soils of the Khorat Plateau. Yields in this environment depend on the timing and amount of monsoon rainfall, on floods along the Mun, Chi and Mekong river systems, on heat and humidity around flowering, and on how farmers allocate land, labour and fertiliser between rice and upland cash crops such as cassava, sugarcane and maize. Timely district-level yield forecasts are therefore valuable for food-security monitoring, for targeting drought and flood relief, for the design and settlement of crop insurance, and for market and price-support policy.

The data available for such forecasts have never been richer. Optical satellite time series describe canopy development (Tucker, 1979; Huete et al., 2002), water-sensitive indices track flooding and transplanting in paddies (Gao, 1996; Xiao et al., 2005), thermal sensors measure land-surface temperature (Wan, 2008), gridded rainfall and land reanalyses provide weather and soil moisture (Funk et al., 2015; Muñoz-Sabater et al., 2021), and cloud platforms make these archives accessible at the scale of administrative units (Gorelick et al., 2017). Machine-learning (ML) models that combine such predictors now routinely report high accuracy (Jeong et al., 2016; Cai et al., 2019; Khaki and Wang, 2019; van Klompenburg et al., 2020).

Three methodological problems nevertheless limit what can be learned from this literature. The first is *evaluation leakage*. Yield data form a spatio-temporal panel in which errors are correlated within seasons and among neighbouring units. Randomly partitioned cross-validation lets a model learn from future seasons and from neighbours in the

same season, and seemingly innocuous predictors—contemporaneous prices, anomalies computed against full-period means, spatial averages of neighbours' outcomes—import information that would not be available at forecast time (Kaufman et al., 2012; Roberts et al., 2017; Kapoor and Narayanan, 2023). The resulting optimism is well documented for time series and spatial ecology (Tashman, 2000; Bergmeir and Benítez, 2012; Roberts et al., 2017) but is rarely quantified for multi-source yield models. The second is *explanation without ground truth*. Feature attributions such as SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020) are increasingly used to rank the drivers of yield, yet with dependent predictors credit can be distributed in counter-intuitive ways (Aas et al., 2021; Hooker et al., 2021), and with real data the true drivers are unknown, so the faithfulness of such explanations cannot be checked. The third is *incomplete integration*. Economic predictors—own and competing-crop prices that govern land and input allocation (Nerlove, 1956)—and explicitly spatial predictors are seldom evaluated alongside weather and satellite data, and point forecasts are rarely accompanied by calibrated uncertainty (Vovk et al., 2005; Lei et al., 2018).

This study addresses these problems with a simulation benchmark. Following the logic of simulation studies for statistical methods (Morris et al., 2019), we construct a process-informed data-generating process (DGP) for main-season rice in 200 districts of Northeast Thailand over 18 seasons. The DGP encodes agronomic mechanisms (water stress at establishment and late in the season, floods, heat, humid-season disease, salinity), economic mechanisms (input intensity, competition from cassava, sugarcane and maize, and a harvest-to-price feedback), persistent management and local pest outbreaks. It emits observations that mimic real sources—monthly weather and soil moisture, cloud-contaminated satellite indices, national and local prices, and imperfect map layers—and it stores every additive yield component, so that model explanations can be compared with the truth. The same code ingests real data with an identical schema, so the benchmark doubles as a reproducible template for empirical work. We stress at the outset that all results reported here concern the synthetic benchmark; they are not empirical statements about Thai rice yields.

We ask four research questions and state hypotheses for each, formulated at the design stage:

RQ1 (accuracy). How much do multi-source learners improve on persistence, climatology and an econometric panel model under leakage-safe evaluation? **H1**: multi-source learners are significantly more accurate than all baselines.

RQ2 (evaluation). How much do random cross-validation and subtle feature leakage overstate skill? **H2**: both produce materially optimistic error estimates.

RQ3 (attribution). Which feature groups carry predictive information, and do SHAP attributions recover the true mechanisms? **H3**: satellite features carry unique information (pest outbreaks invisible to weather data), whereas the moisture, water and satellite groups partly substitute for each other. **H4**: SHAP recovers strong, well-observed mechanisms but not weak or year-confounded ones such as price effects.

RQ4 (timeliness and uncertainty). How does skill evolve during the season, and are prediction intervals calibrated? **H5**: skill increases with the data cut-off month and the value of satellite data grows after panicle initiation. **H6**: conformal intervals are near-nominal on average but under-cover in shock seasons.

The paper makes five contributions: (i) an open, documented benchmark DGP with ground-truth yield components for multi-source yield forecasting; (ii) a quantification of how much common evaluation shortcuts inflate apparent skill; (iii) a ground-truth test of grouped SHAP attributions in the presence of dependent predictor groups; (iv) evidence on the in-season value of satellite data and on the calibration of distribution-free prediction intervals; and (v) a leakage-tested, fully reproducible code base with a documented pathway to real Thai data. Section 2 reviews related work, Section 3 describes the benchmark, Section 4 the methods and Section 5 the results; Section 6 discusses implications and limitations and Section 7 concludes.

2 Related work

Process-based and statistical yield models. Process-based crop models such as DSSAT and APSIM simulate phenology, water and nutrient balances and yield formation from first principles (Jones et al., 2003; Holzworth et al., 2014); they are mechanistically transparent but demand management and soil inputs that are rarely available at district scale. Statistical models relate historical yields to weather and have revealed strongly non-linear temperature responses (Schlenker and Roberts, 2009), quantified the contribution of climate trends, variability and extremes to production (Lobell et al., 2011; Ray et al., 2015; Lesk et al., 2016), and highlighted the joint role of soil moisture and evaporative demand (Rigden et al., 2020). Their strengths and pitfalls are reviewed by Lobell and Burke (2010), and seasonal forecasting applications by Basso and Liu (2019). Our DGP borrows the threshold-type responses of this literature, and our panel fixed-effects baseline follows its econometric tradition (Wooldridge, 2010).

Remote sensing of crop yield. Vegetation indices such as NDVI and EVI summarise canopy greenness and photosynthetic capacity (Tucker, 1979; Huete et al., 2002); water-sensitive indices detect flooding and transplanting of paddy rice (Gao, 1996; Xiao et al., 2005); and land-surface temperature captures canopy stress (Wan, 2008). Regressions on seasonal index metrics underlie operational forecasts (Becker-Reshef et al., 2010; Johnson, 2014), while scalable approaches combine imagery with crop-model simulations (Lobell et al., 2015; Burke and Lobell, 2017). Gridded precipitation (Funk et al., 2015), land reanalysis (Muñoz-Sabater et al., 2021) and cloud computing (Gorelick et al., 2017) make district-level feature extraction routine; monsoon cloud cover, however, leaves gaps that must be handled explicitly.

Machine learning for yield and for tabular data. Random forests (Breiman, 2001) and gradient boosting (Friedman, 2001; Ke et al., 2017) are the workhorses of ML yield studies (Jeong et al., 2016; Cai et al., 2019; van Klompenburg et al., 2020), and deep architectures have been applied to image histograms and weather sequences (You et al., 2017; Khaki and Wang, 2019). On medium-sized tabular data tree ensembles remain hard to beat (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022), which matters at district scale, where a few thousand unit-seasons are typical.

Evaluation and leakage. Out-of-sample evaluation of forecasts must respect temporal order (Tashman, 2000; Hyndman and Athanasopoulos, 2021); ordinary cross-validation is valid for time series only under restrictive conditions (Bergmeir and Benítez, 2012), and spatial, temporal or hierarchical dependence calls for blocked designs (Roberts et al., 2017). Leakage—the use of information unavailable at prediction time—is a pervasive cause of over-optimism in applied ML (Kaufman et al., 2012; Kapoor and Narayanan, 2023). Differences in forecast accuracy should be tested formally, for example with the Diebold–Mariano test (Diebold and Mariano, 1995) and its small-sample correction (Harvey et al., 1997), using resampling schemes that respect dependence (Künsch, 1989; Efron and Tibshirani, 1993).

Explainability with dependent predictors. SHAP values provide additive, axiomatically motivated attributions (Lundberg and Lee, 2017) that can be computed exactly for tree ensembles (Lundberg et al., 2020). When predictors are dependent, however, Shapley-based attributions can spread or shift credit among correlated features and probe the model off the data manifold (Aas et al., 2021; Hooker et al., 2021; Molnar, 2022). Because real yield datasets lack ground truth, the faithfulness of attributions to the data-generating mechanisms has rarely been assessed for crop models; simulation with known truth is the natural remedy (Morris et al., 2019).

Uncertainty quantification. Conformal prediction provides distribution-free predictive intervals under exchangeability (Vovk et al., 2005; Lei et al., 2018), with quantile-based (Romano et al., 2019) and adaptive variants for distribution shift (Gibbs and Candès, 2021). Seasonal yield data are not exchangeable—shock seasons cluster errors—so the calibration of conformal intervals in this setting is an empirical question that we examine directly.

3 Study region and benchmark design

3.1 Study region

Northeast Thailand (Isan) comprises 20 provinces on the Khorat Plateau, bounded by the Mekong River to the north and east, the Phetchabun range to the west and the Dangrek range to the south (Fig. 1a,b). The plateau is drained by the Chi and Mun rivers, which join before flowing into the Mekong. Most rice in the region is grown as a single rainfed wet-season crop on light-textured, often sandy soils, parts of which are affected by salinity, and upland fields are increasingly planted with cassava and sugarcane. The monsoon delivers most rainfall between May and October, with a drier, rain-shadowed south-west and a wetter north-east, so that early-season dry spells, an early end of the rains and floods along the river systems dominate the year-to-year variability of yields.

The benchmark reproduces these features qualitatively (Sections 3.4–3.8) but contains no observed data from the region. To relate the simulation to real geography, each simulated district is assigned to the province of its nearest provincial capital and district values are averaged by province. Fig. 1c–f shows the simulated satellite signal at this level for four seasons. In 2019 and 2023, the two shock seasons of the test period, peak-season greenness was below normal in 15 and 16 of the 20 provinces (regional means -0.020 and -0.020), whereas in 2021 and 2025 it was above normal in 16 and 16 provinces (regional means $+0.022$ and $+0.018$). The spatial patterns differ between seasons—the 2019 deficit is concentrated in the north-west of the region and the 2023 deficit in the west and south-west—reflecting

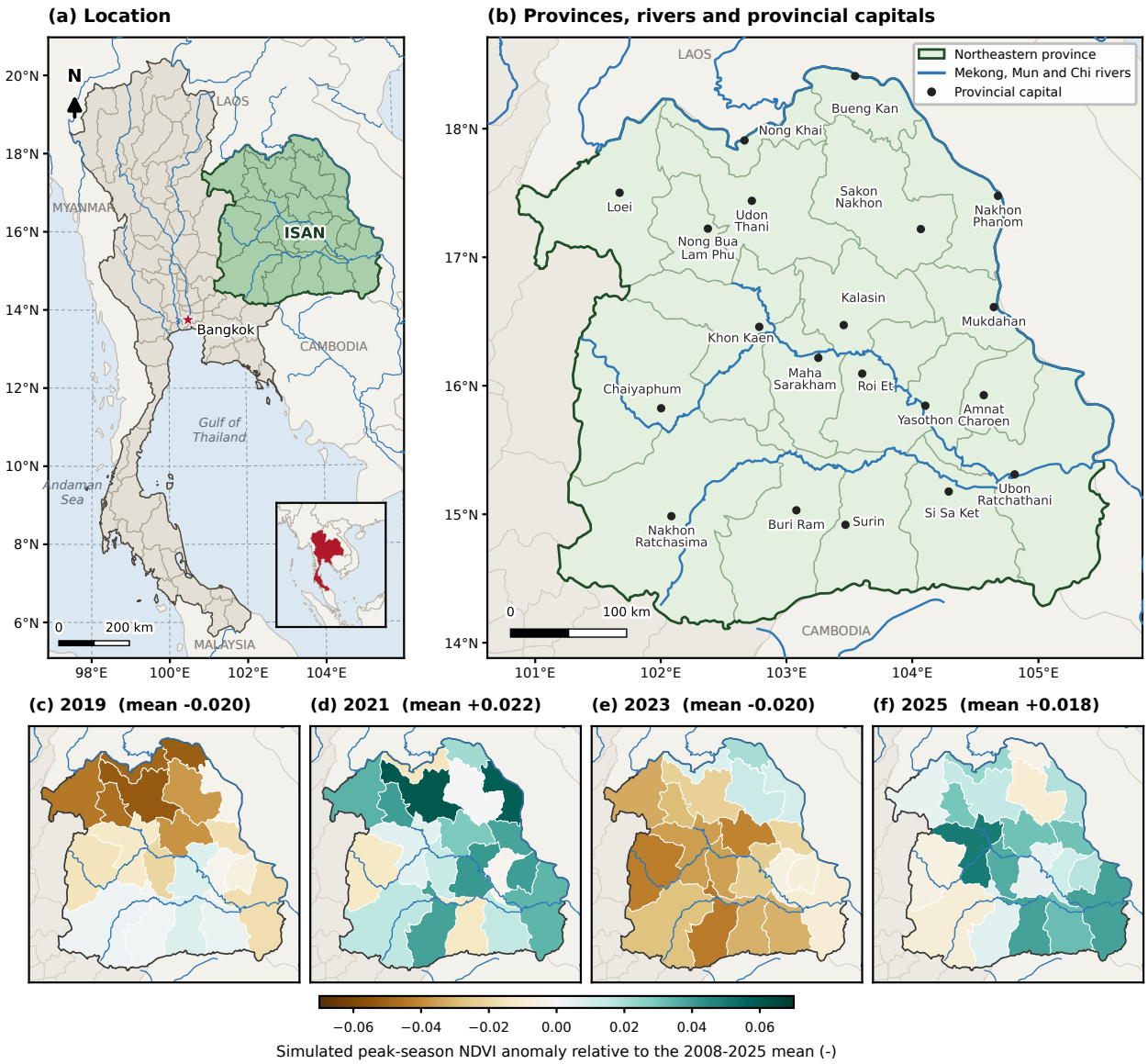


Figure 1: Study region. (a) Northeast Thailand (Isan; green) within Thailand and Southeast Asia (inset); star: Bangkok. (b) The 20 northeastern provinces with the Mekong, Mun and Chi rivers and the provincial capitals. (c–f) Simulated peak-season NDVI anomaly, relative to each simulated district's 2008–2025 mean and averaged over the simulated districts of each province, in four seasons (panel titles give the regional mean); these are benchmark values, not satellite observations. Boundaries and rivers: Natural Earth (Natural Earth, 2026); projection WGS 84 / UTM zone 47N (EPSG:32647).

the spatially correlated rainfall anomalies of the data-generating process; a model that is to be useful must therefore resolve within-region contrasts, not only the regional mean.

3.2 Forecasting problem

The unit of analysis is a district (*amphoe*) i in season t; the target y_{it} is the main-season rice yield in kg rai^{-1} (1 rai $= 0.16$ ha, hence 1 kg rai^{-1} $= 6.25$ kg ha^{-1}). The season runs from sowing in May to harvest in November and is divided into establishment (May–July), vegetative (August–September) and reproductive (October–November) stages. A forecast for season t issued at the end of month τ may use any information available by then: yields of earlier seasons, weather and satellite observations up to τ, prices observed before sowing, and static map layers.

3.3 Design of the simulation study

We follow the aims–data-generating mechanisms–estimands–methods–performance structure recommended for simulation studies (Morris et al., 2019). The aims are RQ1–RQ4. The data-generating mechanisms are described below, in Table 1 and in Appendix A. The estimands are out-of-sample forecast errors on the test seasons 2018–2025 and, for RQ3, the true additive yield components. The methods are the learners and protocols of Section 4, and performance is measured by the error, coverage and recovery statistics of Section 4.4. The main world uses seed 42; ten further worlds and four scenario worlds quantify Monte-Carlo and structural variability (Section 4.9).

3.4 Spatial layers

District centroids are placed inside an outline of the region with Mitchell's best-candidate algorithm, which yields an even, administrative-like spacing; Voronoi cells are used only for display (Fig. 2). Static layers combine smooth deterministic patterns with spatially correlated Gaussian random fields (kernel-smoothed white noise with length scales of 40–150 km): elevation with the Phetchabun, Dangrek and Phu Phan uplands and lower river valleys; slope; soil sand content, soil organic carbon (SOC) and pH; salinity hot spots in the central lowlands; a four-class land-suitability index (S1, S2, S3, N); irrigated and paddy shares; flood frequency, which decays with distance from the Mekong, Mun and Chi; and distances to provincial towns (markets) and sugar mills. The observed sand, SOC, salinity and irrigation layers equal the true values plus measurement error, mimicking coarse soil maps and outdated irrigation inventories.

3.5 Weather, water and soil moisture

A season-level, ENSO-like index e_t (positive: dry and hot) loosely follows the historical phase sequence (Ropelewski and Halpert, 1987). Monthly rainfall is log-normal around a district rainfall normal that increases from the rain-shadowed south-west to the wet north-east, with seasonal, monthly and idiosyncratic spatially correlated anomalies; dry phases depress June–July rainfall most. Temperature follows a lapse rate, a warming trend, e_t and rainfall anomalies, and relative humidity rises with rainfall. Soil moisture follows a texture-dependent bucket model, so that identical rainfall leaves sandy soils drier. Reservoir storage at sowing in each of three basins depends on the previous season's dryness and supplies water to irrigated areas.

3.6 Prices and economic behaviour

National farm-gate prices of rice, cassava, sugarcane, maize and urea follow correlated log-AR(1) processes with occasional shocks such as fertiliser price spikes. Local prices subtract transport costs proportional to the distance to a market or, for cane, to a sugar mill. Two behavioural channels act on yields through lag-1 prices known at sowing: a higher rice-to-urea price ratio raises input intensity, and a higher revenue of competing crops relative to rice diverts labour and inputs from rice, more strongly in non-irrigated districts. A harvest-to-price feedback lowers the rice price in seasons with a large regional harvest, so that same-season prices carry information about the season's yields—a realistic source of endogeneity that the leakage audit exploits.

3.7 Yield equation

Yield is the clipped sum of additive components,

$$y_{it} = \min\Big\{850,\ \max\big\{120,\ b_i + \tau\,(t-2008) + D_{it} + F_{it} + H_{it} + R_{it} + S_{it} + P_{it} + m_{it} + \pi_{it} + \varepsilon_{it}\big\}\Big\}, \tag{1}$$

where b_i is a district base level, D drought (establishment and late season), F flood, H heat, R humid-season disease, S salinity aggravated by drought, P price effects, m an AR(1) management effect, π local pest outbreaks and ε unobservable noise (Table 1). All stress terms are hinge functions, which creates non-linearity, and several are interactions (irrigation $\times$ reservoir buffering, salinity $\times$ drought, competition $\times$ irrigation). Pest outbreaks occur in 2016, 2022, 2023 and affect districts within roughly 70 km of a random centre; they are visible to the satellite indices but to no weather variable.

Table 1: Additive yield mechanisms of the benchmark DGP (contributions in kg rai^{-1}), parameter values used in the base world, and the feature groups that observe each mechanism. q_i: soil-quality index; I_i: irrigated share; $\bar{s}$: true soil moisture (JJ: Jun–Jul, SO: Sep–Oct); S_{it}: irrigation supply (irrigated share × reservoir storage); f_i: flood frequency; w_{it}: Aug–Sep rainfall relative to normal; σ_i: saline share; $\tilde{c}_{it}$, $\tilde{p}_{it}$: standardised competing-revenue and rice/urea indices from lag-1 local prices; $\bar{a}_t$: regional mean yield anomaly. Scenario variants multiply all threshold terms by 1.8 and add a compound heat–drought penalty (non-linear), triple map-layer measurement error, raise cloud frequency 1.7-fold, or remove pests.

Mechanism	Contribution to y_{it}	Parameters	Observed by
Base level	$b_i = \beta_0 + \beta_q q_i + \beta_I I_i + \sigma_u u_i + \sigma_\eta \eta_i$	$\beta_0 = 365$, $\beta_q = 30$, $\beta_I = 115$, $\sigma_u = 18$, $\sigma_\eta = 12$	Geography, map, history
Technology trend	$\tau\,(t - 2008)$	$\tau = 2.2$	History
Establishment drought	$-\gamma_E \max\{0, \theta_E - (\bar{s}_{it}^{JJ} + \kappa_E S_{it})\}$	$\gamma_E = 1100$, $\theta_E = 0.24$, $\kappa_E = 0.12$	Moisture, water, satellite
Late-season drought	$-\gamma_L \max\{0, \theta_L - (\bar{s}_{it}^{SO} + \kappa_L S_{it})\}$	$\gamma_L = 900$, $\theta_L = 0.22$, $\kappa_L = 0.1$	Moisture, water, satellite
Flood	$-\min\{c_F,\ \gamma_F (f_i/0.3) \max(0, w_{it} - \theta_F)\}$	$\gamma_F = 160$, $\theta_F = 1.3$, $c_F = 150$	Water, map, satellite
Heat at flowering	$-\gamma_H \sum_{m\in\{9,10\}} \max(0, T_{itm}^{\max} - \theta_H)$	$\gamma_H = 6$, $\theta_H = 34\,^\circ$C	Climate
Humid-season disease	$-\gamma_D \max(0, \overline{RH}_{it}^{AS} - \theta_D)$	$\gamma_D = 4$, $\theta_D = 84\,\%$	Moisture
Salinity × drought	$-\gamma_S \sigma_i \max(0, \theta_S - \bar{s}_{it}^{JJ})$	$\gamma_S = 250$, $\theta_S = 0.26$	Geography, moisture
Prices & competition	$-\gamma_C \tilde{c}_{it}(1 - I_i) + \gamma_P \tilde{p}_{it}(0.6 + I_i)$	$\gamma_C = 12$, $\gamma_P = 10$	Prices, map
Harvest-price feedback	$\ln P_t^{\text{rice}} \leftarrow \ln P_t^{\text{rice}} - \kappa \bar{a}_t/100$	$\kappa = 0.25$	(price endogeneity)
Management	$m_{it} = \rho\, m_{i,t-1} + \sigma_m \xi_{it}$	$\rho = 0.6$, $\sigma_m = 14$	History, satellite
Pest outbreaks	$-A \exp\{-d_{ic}^2/(2r^2)\}$ in 3 seasons	$A = 75$, $r = 70$ km	Satellite only
Noise	$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$	$\sigma_\varepsilon = 16$	None (irreducible)

3.8 Satellite observation model

Monthly NDVI follows a bell-shaped seasonal curve whose amplitude saturates in the *visible* part of yield (all components except ε) and whose peak is delayed when early-season soil moisture is low; flooding depresses NDVI in September–October. EVI responds less saturatingly, LSWI tracks soil moisture, transplanting water and floods, and LST rises with maximum temperature and dryness and falls with greenness. Monsoon clouds remove or contaminate monthly composites with month-specific probabilities of up to 50% in August; contamination lowers the vegetation indices and LST, as in real composites.

3.9 Ground truth and scenarios

For every district-season the simulator stores the true components of Eq. (1). Four scenario variants of the main world test structural robustness: (a) stronger non-linearity (all threshold terms multiplied by 1.8 plus a compound heat–drought penalty), (b) three-fold map-layer measurement error, (c) 1.7-fold cloud contamination, and (d) no pest outbreaks, that is, no satellite-only signal.

4 Methods

4.1 Features

From the raw tables we construct 81 features in eight groups (Table 2). Weather and satellite variables are aggregated by growth stage (sums for rain; means, minima and maxima for moisture, temperature and indices) and complemented by season metrics: total rain and its anomaly relative to the district's previous seasons, the number of dry and hot months, growing degree-days (base 10 °C), vapour-pressure deficit, and the NDVI peak, area, timing and green-up rate after linear gap-filling of cloud-missing months, together with a clear-sky fraction. Price features convert lag-1 national prices into local farm-gate prices; price ratios and a log competing-revenue index summarise relative profitability. Map features include distances, irrigated and paddy shares, flood frequency and k-nearest-neighbour ($k = 8$) spatial lags of the neighbours' previous-season yields and current-season NDVI anomalies.

Four rules make the features leakage-safe: (i) yield history uses previous seasons only; (ii) prices are lagged one season; (iii) weather and satellite data are truncated at the forecast cut-off month; and (iv) anomalies are computed against each district's previous seasons. The rules are enforced by automated tests that perturb all data of later seasons and verify that the features of earlier seasons are unchanged, and—as a check of the test itself—that each deliberately leaky variant of Section 4.5 fails it.

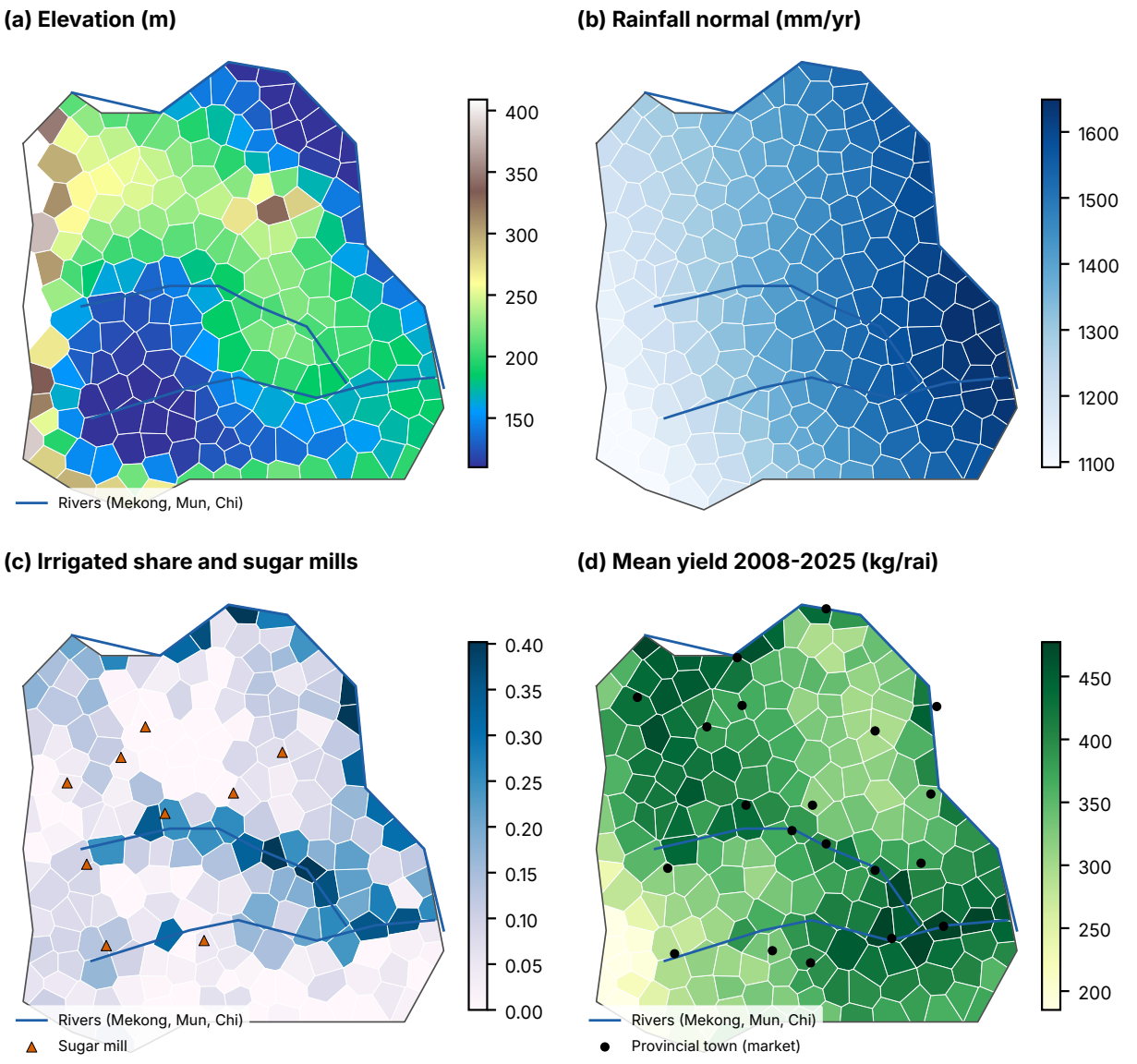


Figure 2: The synthetic study region (main world, 200 districts; Voronoi cells for display only). (a) Elevation; (b) rainfall normal; (c) observed irrigated share and sugar mills; (d) mean simulated yield 2008–2025 and the provincial towns used as markets. Blue lines: simplified Mekong, Mun and Chi rivers.

4.2 Learners

We compare eight learners of increasing flexibility. *Persistence* forecasts last season's yield and *climatology* the mean of the previous three seasons, $\bar{y}_{it}^{(3)}$. The *panel fixed-effects* (FE) model is the econometric baseline: a within estimator with district effects, a linear trend and all time-varying covariates, with lagged yields excluded to avoid dynamic-panel bias and a small ridge penalty for numerical stability (Wooldridge, 2010). *Ridge regression* (Hoerl and Kennard, 1970) uses standardised, median-imputed features with the penalty chosen by efficient leave-one-out cross-validation within the training data. The *random forest* (Breiman, 2001) has 200 trees. The *multilayer perceptron* (MLP; hidden layers of 64 and 32 units, early stopping) represents neural approaches. *LightGBM* (Ke et al., 2017) is a histogram-based gradient-boosting machine that handles missing values natively. The *ensemble* averages LightGBM and Ridge with equal weights, combining the threshold-and-interaction capacity of trees with the smooth extrapolation of a linear model; fixed weights avoid an additional layer of fitting (cf. Wolpert, 1992).

Table 2: Feature groups, number of features, content, and the real-data sources that can replace the synthetic inputs without code changes (OAE: Office of Agricultural Economics; DOAE: Department of Agricultural Extension; TMD: Thai Meteorological Department; RID: Royal Irrigation Department; OCSB: Office of the Cane and Sugar Board; LDD: Land Development Department; GISTDA: Geo-Informatics and Space Technology Development Agency).

Group	n	Features	Real-data sources
Previous yields	5	lags $t-1$, $t-2$; previous-3-season mean and s.d.; 5-season trend	OAE / DOAE statistics
Humidity & soil moisture	12	RH, VPD, soil-moisture mean and minimum per growth stage	ERA5-Land; SMAP
Water (rain, reservoir, irrigation)	8	rain per stage and season; anomaly vs. past; dry months; reservoir storage; irrigation supply	CHIRPS; TMD; RID reservoirs
Climate (temperature, GDD)	9	T_{mean}, $T_{\max}$ per stage; GDD; hot months; rainfall normal	ERA5-Land; TMD normals
Satellite (NDVI, EVI, LSWI, LST)	19	NDVI/EVI maxima, LSWI/LST means per stage; NDVI peak, area, timing, green-up; clear-sky fraction; anomalies	MODIS MOD13Q1, MOD09A1, MOD11A2; Sentinel-1/2
Prices (own & competing crops)	11	lag-1 local prices of rice, cassava, sugarcane, maize; ratios; competing-revenue index; rice/urea ratio	OAE farm-gate prices; OCSB cane price
Geography (terrain, soil)	9	elevation, slope, sand, SOC, pH, saline share, suitability class, coordinates	SRTM; LDD soil maps; SoilGrids
Map layers & spatial lags	8	distances to river, mill and market; irrigated and paddy share; flood frequency; k-NN spatial lags	RID; LDD land use; GISTDA floods; OSM
Total	81		

Two target formulations are considered: the yield level y_{it}, and the anomaly $y_{it} - \bar{y}_{it}^{(3)}$, which removes district effects and slow trends that trees cannot extrapolate; anomaly forecasts are added back to $\bar{y}_{it}^{(3)}$.

4.3 Validation protocol

All evaluation uses a rolling origin (Tashman, 2000): the forecast for season t comes from a model fitted on seasons $s < t$ only. Seasons 2008–2010 serve only to initialise lagged features. LightGBM hyper-parameters and the target formulation are tuned jointly by random search (15 configurations $\times$ 2 targets) on the validation seasons 2015–2017, each forecast with rolling origin, and the main model is then selected among LightGBM, Ridge and the ensemble on the same validation seasons. The test seasons 2018–2025 (8 seasons $\times$ 200 districts = 1,600 forecasts) were used only after all choices had been frozen.

4.4 Performance measures and inference

We report RMSE, MAE, MAPE and the coefficient of determination R^2, together with two skill measures that are not inflated by the large between-district variance:

$$R^2_{\text{anom}} = 1 - \frac{\sum_{it}(y_{it} - \hat{y}_{it})^2}{\sum_{it}(a_{it} - \bar{a})^2}, \qquad \text{Skill} = 1 - \frac{\text{MSE}}{\text{MSE}_{\text{clim}}}, \qquad a_{it} = y_{it} - \bar{y}_{it}^{(3)}. \tag{2}$$

R^2_{anom} is the share of the deviation from the recent mean that is predicted, and Skill is the reduction of the climatology's mean squared error. Because errors are strongly correlated within seasons, uncertainty is quantified with a season-block bootstrap that resamples whole test seasons (Künsch, 1989; Efron and Tibshirani, 1993) ($B = 2{,}000$). Accuracy differences relative to the main model are tested with the Diebold–Mariano test (Diebold and Mariano, 1995) applied to the season-level differentials d_t of mean squared errors, with the correction of Harvey et al. (1997) for $T = 8$ one-step-ahead forecasts,

$$\text{DM}^* = \sqrt{\frac{T-1}{T}}\,\frac{\bar{d}}{\sqrt{\hat{\gamma}_0/T}} \;\sim\; t_{T-1}, \tag{3}$$

where $\hat{\gamma}_0$ is the sample variance of d_t; p-values are adjusted with the Holm procedure (Holm, 1979).

4.5 Leakage audit (E2)

LightGBM, Ridge and the ensemble are re-estimated under four flawed protocols and scored on the same 1,600 test forecasts: (i) random 5-fold cross-validation over all district-seasons, which trains on future seasons and same-season neighbours; (ii) same-season instead of lag-1 prices; (iii) an additional spatial lag of neighbours' current-season yields; and (iv) weather and NDVI anomalies computed against the full-period district mean. Optimism is expressed as the relative reduction of RMSE compared with the leakage-safe protocol.

4.6 Attribution and ground-truth recovery (E3–E4)

Predictive value is measured by drop-one-group ablation: each group is removed, the model is re-trained in every fold and the RMSE change is computed with season-bootstrap intervals. When the history group is removed, the level target is used, because the anomaly target presupposes past yields. Attributions are computed for every test forecast: TreeSHAP for LightGBM (Lundberg et al., 2020) and exact linear SHAP, $\phi_j = \beta_j(x_j - \mathbb{E}x_j)$, for Ridge; because Shapley values are additive in the model output, the ensemble's attribution is the average of the two. Group attributions are sums, $\Phi_G = \sum_{j \in G} \phi_j$. Ground-truth recovery is assessed by the Pearson correlation between each true component of Eq. (1) and Φ_G for every group, and by comparing binned SHAP dependence curves with the binned true component along the main observed driver.

4.7 In-season forecasting (E5)

Features are rebuilt with data cut-offs at the end of July, August, September, October and November, and the full protocol is repeated with and without the satellite group. The difference measures the marginal value of satellite data at each forecast date.

4.8 Prediction intervals (E6)

Prediction intervals use rolling split-conformal calibration: for test season t, the interval is $[\hat{y}_{it} + q_{\alpha/2},\ \hat{y}_{it} + q_{1-\alpha/2}]$, where q are empirical quantiles of the main model's out-of-sample errors in all earlier seasons from 2014 onwards. This respects the forecasting information set. We evaluate 80% and 90% intervals overall, by season, and separately in *shock seasons*, defined as seasons whose regional mean anomaly $\bar{a}_t$ lies more than one standard deviation below the mean of all seasons.

4.9 Robustness (E7)

The tuned configuration is applied without re-tuning to ten independently simulated replicate worlds (seeds 1000–1009) and to the four scenario worlds. For each world we record the RMSE of climatology, Ridge, LightGBM and the ensemble, of the ensemble without satellite features, and of the ensemble scored by random cross-validation.

4.10 Implementation and reproducibility

The pipeline is written in Python 3.12 with NumPy, pandas, scikit-learn (Pedregosa et al., 2011) and LightGBM 4.7, and the complete set of experiments runs on a single CPU core in about 7 minutes. Every number in this paper is written to a macro file by code from saved results. Raw inputs follow a documented schema; a Google Earth Engine script extracts their real counterparts (MODIS vegetation indices, LSWI and LST, CHIRPS rainfall, ERA5-Land weather and soil moisture, SRTM terrain and a cropland mask) for any set of district polygons (Table 2). Maps are drawn with GeoPandas from Natural Earth boundaries (Natural Earth, 2026) by `scripts/thailand_map.py`.

5 Results

5.1 Accuracy of multi-source forecasts (RQ1)

Table 3 and Fig. 4 summarise the 1,600 test forecasts. Climatology reached an RMSE of 47.1 kg rai^{-1} and persistence 59.2 kg rai^{-1}. Every multi-source learner was more accurate. LightGBM had the lowest test error (27.4 kg rai^{-1}, i.e. 0.171 t ha^{-1}; MAPE 6.4%), 42% below climatology, followed by the random forest (28.2) and the validation-selected ensemble (29.9 kg rai^{-1}; 95% CI 24.7–37.4), which reduced the climatology error by 36% (difference 17.1 kg rai^{-1}, CI 3.6–27.5). The multilayer perceptron (32.2), the panel fixed-effects model (35.2) and Ridge regression (37.9) were less accurate. The skill measures confirm that the gains concern the season-to-season deviations that matter for forecasting: LightGBM explained a share of 0.66 of the variance of the anomaly relative to the previous-three-season mean (Skill 0.66), the ensemble 0.59 (Skill 0.60).

Formal tests were less conclusive than these magnitudes suggest. With only eight test seasons and Holm adjustment over eight comparisons, no Diebold–Mariano test against the ensemble was significant at the 5% level (smallest

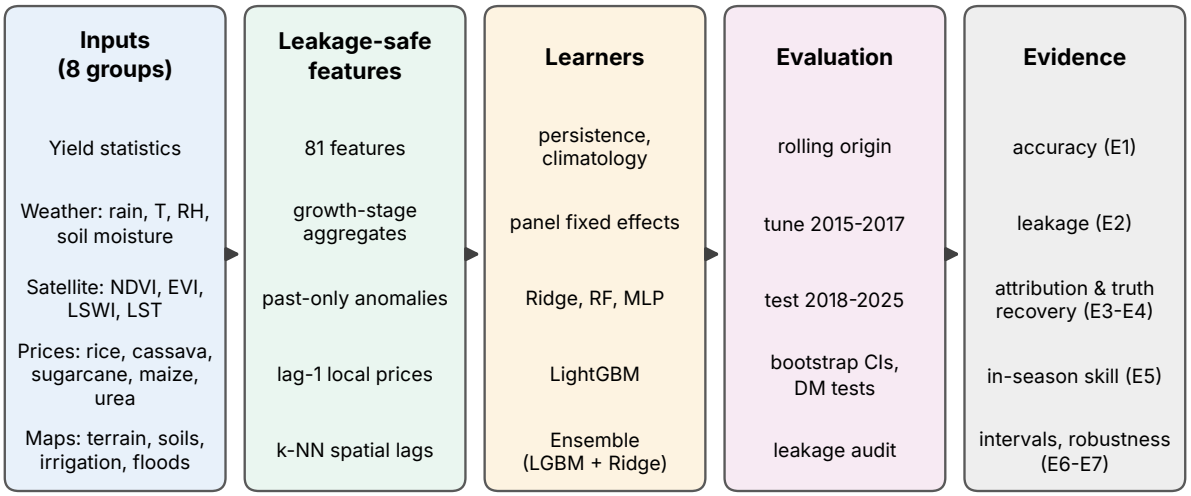


Figure 3: Study workflow: inputs in eight groups are turned into leakage-safe features, eight learners are evaluated with a rolling origin, and seven experiments (E1–E7) provide the evidence for RQ1–RQ4. The simulator's stored true yield components enable the ground-truth test of explanations (E4).

Table 3: Out-of-sample accuracy on the test seasons 2018–2025 (rolling origin; 1,600 district-season forecasts). RMSE and MAE in kg rai^{-1} (multiply by 0.00625 for t ha^{-1}); brackets: 95% season-block bootstrap CI. R^2_{anom}: skill on deviations from the previous-3-season mean; Skill: $1 - \text{MSE}/\text{MSE}_{\text{clim}}$. p_{DM}: Holm-adjusted Diebold–Mariano (HLN) test against the main model (*).

Model	RMSE [95% CI]	MAE	MAPE (%)	R^2	R^2_{anom}	Skill	p_{DM}
LightGBM	27.4 [26.1, 29.0]	21.7	6.4	0.89	0.66	+0.66	1.000
Random forest	28.2 [26.8, 29.6]	22.2	6.6	0.88	0.64	+0.64	1.000
Ensemble (LightGBM + Ridge) *	29.9 [24.7, 37.4]	23.4	7.0	0.87	0.59	+0.60	–
LightGBM (alternative target)	31.2 [29.0, 33.4]	24.6	7.3	0.85	0.56	+0.56	1.000
Multilayer perceptron	32.2 [30.3, 34.0]	25.6	7.6	0.84	0.53	+0.53	1.000
Panel fixed effects	35.2 [26.8, 46.5]	27.5	7.9	0.81	0.44	+0.44	1.000
Ridge regression	37.9 [25.4, 54.3]	27.9	8.3	0.78	0.35	+0.35	1.000
Climatology (3-season mean)	47.1 [40.6, 53.2]	37.8	11.4	0.67	-0.00	+0.00	0.251
Persistence	59.2 [49.3, 69.1]	48.1	14.6	0.47	-0.59	-0.58	0.080

adjusted $p = 0.080$, for persistence; $p = 0.251$ for climatology). The season-bootstrap interval of the ensemble's advantage over climatology nevertheless excludes zero, and the replicate worlds (Section 5.7) show that the advantage is systematic.

Model selection on the validation seasons favoured the ensemble (validation RMSE 26.2, against 27.4 for Ridge and 28.1 for LightGBM), but on the test seasons LightGBM was more accurate (RMSE difference -2.5 kg rai^{-1}, 95% CI -9.6 to 2.1). The error by season (Fig. 5) shows why: the linear learners failed in single seasons—Ridge in 2018 (RMSE 81.7) and the panel model in 2020 (65.4)—and the ensemble inherited half of Ridge's failure (50.9 in 2018), whereas LightGBM never exceeded 31.9 kg rai^{-1} in any season. In the remaining seasons the ensemble was usually the most accurate learner, including the shock seasons 2019, 2023, in which climatology and persistence were worst. The level target was preferred to the anomaly target during tuning, and the anomaly-target LightGBM was also less accurate on the test seasons (31.2).

5.2 Leakage inflates apparent skill and hides failures (RQ2)

Table 4 and Fig. 6 compare the leakage-safe estimates with four flawed protocols. For LightGBM—the learner whose honest performance was stable—random 5-fold cross-validation understated the RMSE by 10%, and adding the neighbours' current-season yields by 5%. Anomalies computed against the full-period mean had a negligible effect (1%),

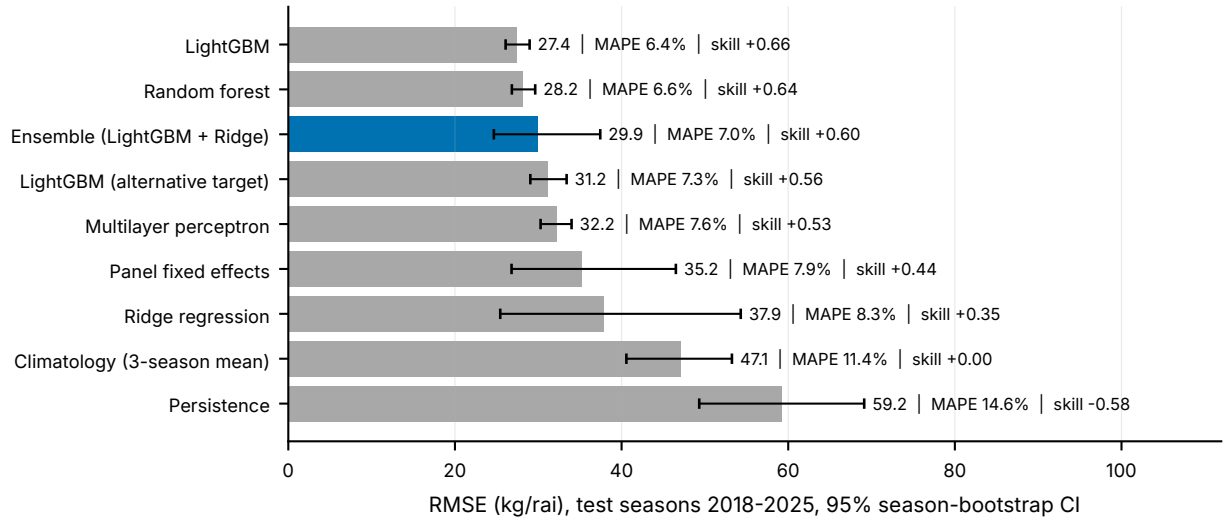


Figure 4: Test RMSE of the eight learners (rolling origin, 2018–2025) with 95% season-block bootstrap intervals; the validation-selected main model is highlighted. Text: RMSE, MAPE and skill relative to climatology.

Table 4: Leakage audit: RMSE (kg rai^{-1}) reported on the same 1,600 test forecasts under the leakage-safe protocol and under four flawed protocols; parentheses: change relative to the leakage-safe RMSE (negative = optimistic bias).

Protocol	LightGBM	Ridge	Ensemble
Rolling origin (leakage-safe)	27.4	37.9	29.9
Random 5-fold CV	24.6 (-10%)	24.7 (-35%)	23.8 (-21%)
Same-season prices	28.0 (+2%)	28.3 (-25%)	25.9 (-14%)
Current neighbour yields	25.9 (-5%)	24.6 (-35%)	24.3 (-19%)
Anomalies vs. full-period mean	27.2 (-1%)	34.4 (-9%)	28.5 (-5%)

and same-season prices did not help (-2%) although they carry information about the current harvest through the price feedback; a plausible explanation is that, with fewer than fifteen training seasons, a predictor that varies mainly between seasons is hard to exploit. Apparent gains were much larger for Ridge and the ensemble (random cross-validation: 35% and 21%), but for a different reason: under the flawed protocols the linear learner no longer had to extrapolate to an unseen season, and its catastrophic failure disappeared. Random cross-validation therefore did not merely inflate skill; it concealed the failure mode altogether. The replicate worlds confirm the pattern: the randomly cross-validated ensemble reported 23.5 ± 0.4 kg rai^{-1} in every world (range 23.1–24.4), whereas its honest RMSE ranged from 24.4 to 55.2.

5.3 Which predictors matter (RQ3)

Drop-one-group ablation (Fig. 7a; Table B.3) identified the satellite group as the only group whose removal materially degraded LightGBM (+25.0%); removing any other group changed its RMSE by between -2.8% and $+1.6\%$. The weather, soil-moisture, price, terrain and history groups are therefore individually substitutable: much of their information is also contained in other groups—soil moisture in rainfall and LSWI, the district level in terrain, soils and neighbours' yields. For the ensemble, removing the satellite group raised the RMSE by +90.3% and removing geography by +15.1%, but removing the moisture, climate or price features *lowered* it (-14.1%, -13.2% and -13.9%), because every reduction of the season-level feature set reduced Ridge's scope for extrapolation. Ablation results of a fragile learner thus mix information content with model stability.

Grouped SHAP shares differed sharply between learners (Fig. 7b). LightGBM attributed 36.4% of the mean absolute SHAP value to the satellite group, 20.1% to history and 11.0% to moisture, but only 6.9% to prices. The ensemble attributed the largest share to prices (28.7%), followed by the satellite (21.6%) and moisture (20.6%) groups. Ablation and SHAP answer different questions—the marginal predictive value of a group versus its use by a particular

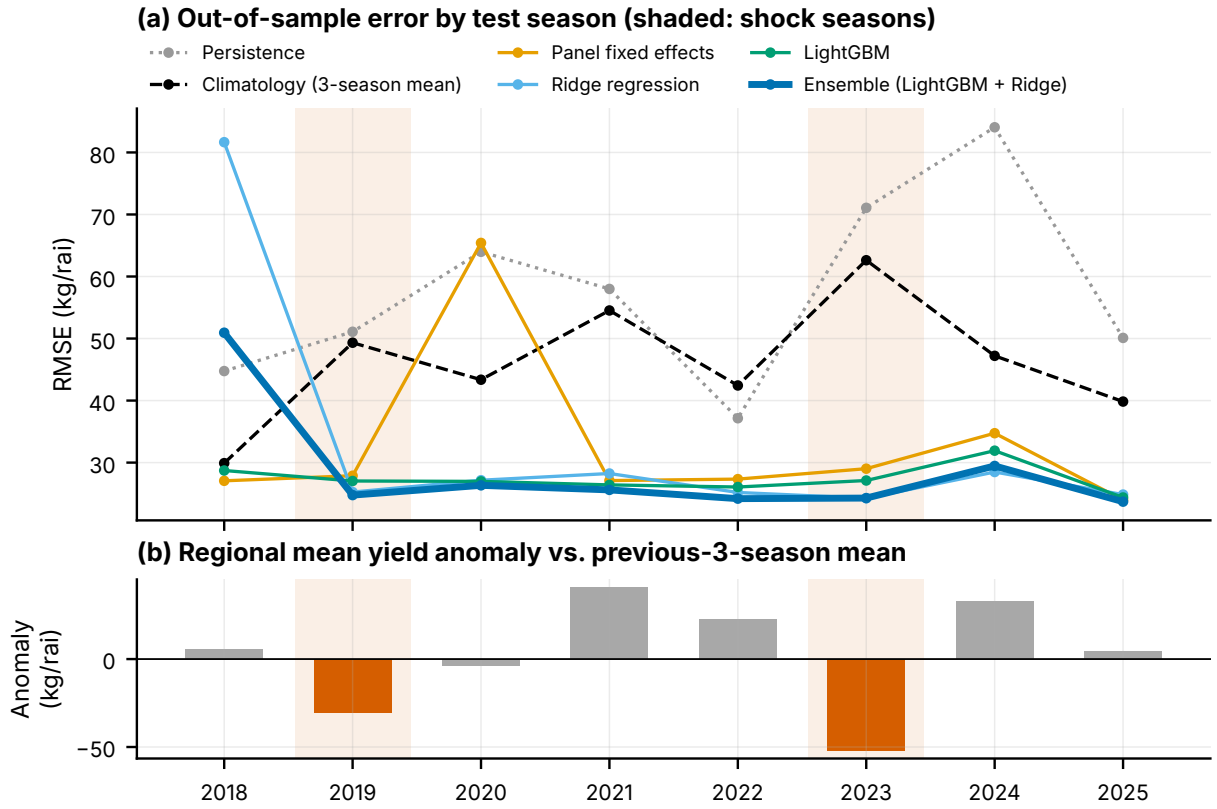


Figure 5: (a) RMSE by test season. The linear learners fail in single seasons (Ridge in 2018, panel fixed effects in 2020), and the ensemble inherits part of Ridge's failure. (b) Regional mean yield anomaly relative to the previous-three-season mean; shaded seasons are shock seasons (anomaly more than one standard deviation below the mean of all seasons).

model—and the ensemble's large price share reflects Ridge's use of season-level price variables rather than their predictive value, since removing them improved the ensemble. The satellite group's value is not restricted to pests: in the world without pest outbreaks the ensemble's RMSE still rose from 25.0 to 43.6 kg rai^{-1} without satellite features, because canopy indices integrate the visible state of the crop, including drought, management and soil effects.

5.4 Do explanations recover the true mechanisms? (RQ3)

Because the simulator stores the true yield components, attributions can be checked directly (Fig. 8, Table 5). Two mechanisms were recovered well: the district base level, through the history, geography and map groups ($r = 0.90$ for the ensemble and 0.92 for LightGBM), and drought, through the moisture, water and satellite groups ($r = 0.84$ and 0.88); the binned SHAP curve of the moisture and water groups reproduced the threshold response of the true drought component (Fig. 9a). Recovery failed in three characteristic ways. First, *substitution among correlated mechanisms*: heat stress around flowering, which co-occurs with dry conditions, was credited mainly to the moisture and water groups, while the climate group's SHAP was negatively correlated with the true heat penalty ($r = -0.39$; Fig. 9d). Second, *absorption of season-level effects*: the technology trend, a pure season effect, was not recovered by the history group ($r = 0.06$) but correlated strongly and negatively with the price group's SHAP (Fig. 8); a predictor that varies mainly between seasons can stand in for any season effect. Third, *weak or diluted signals*: floods ($r = -0.04$), humid-season disease ($r = -0.25$) and management ($r = 0.14$) were essentially not recovered, and pest outbreaks only weakly ($r = 0.16$), partly because their satellite signal is shared with the neighbours' NDVI anomalies in the map group. Learner stability mattered most for prices: LightGBM's price and map attributions tracked the true price effect ($r = 0.52$), whereas the ensemble's did not ($r = -0.02$), and its price SHAP even ran against the true response (Fig. 9c). Unobservable noise was, correctly, attributed to no group.

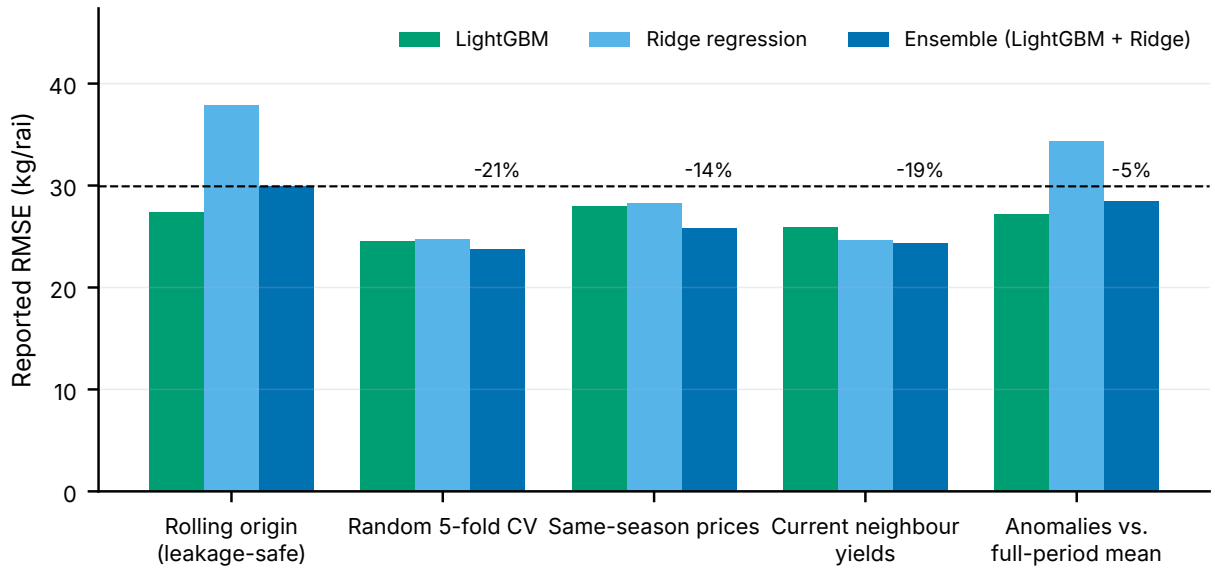


Figure 6: Reported test RMSE under the leakage-safe rolling origin (dashed line: main model) and under four flawed protocols, all scored on the same 1,600 forecasts; labels give the change for the main model.

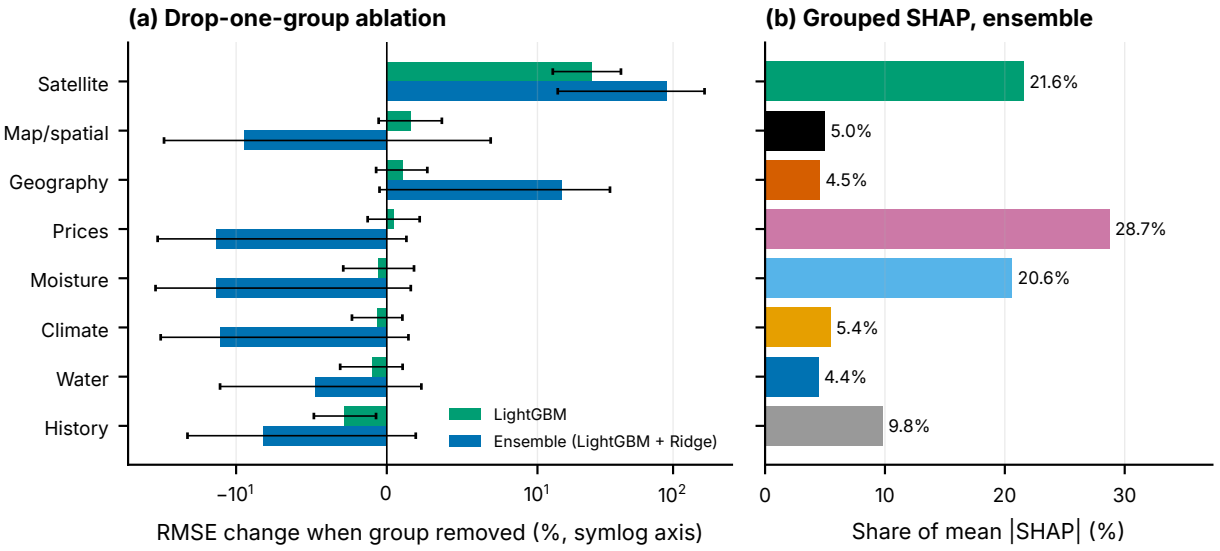


Figure 7: (a) Change in test RMSE when a feature group is removed and the learner re-trained (95% season-bootstrap intervals; symmetric-log axis). (b) Share of mean absolute SHAP values by group for the main model.

5.5 In-season forecasting (RQ4)

LightGBM's RMSE fell from 35.9 kg rai^{-1} with data to the end of July to 35.3 (August), 34.0 (September), 28.5 (October) and 27.4 (November), and it was below climatology (47.1) at every forecast date (Fig. 10a). The value of the satellite group was negligible until September (0.1, 0.4 and 0.9 kg rai^{-1}) and became substantial once the canopy approached its peak (5.8 in October and 6.8 in November; Fig. 10b). Ridge, in contrast, was unusable early in the season (141.0 kg rai^{-1} in July), and the ensemble inherited this failure (77.3).

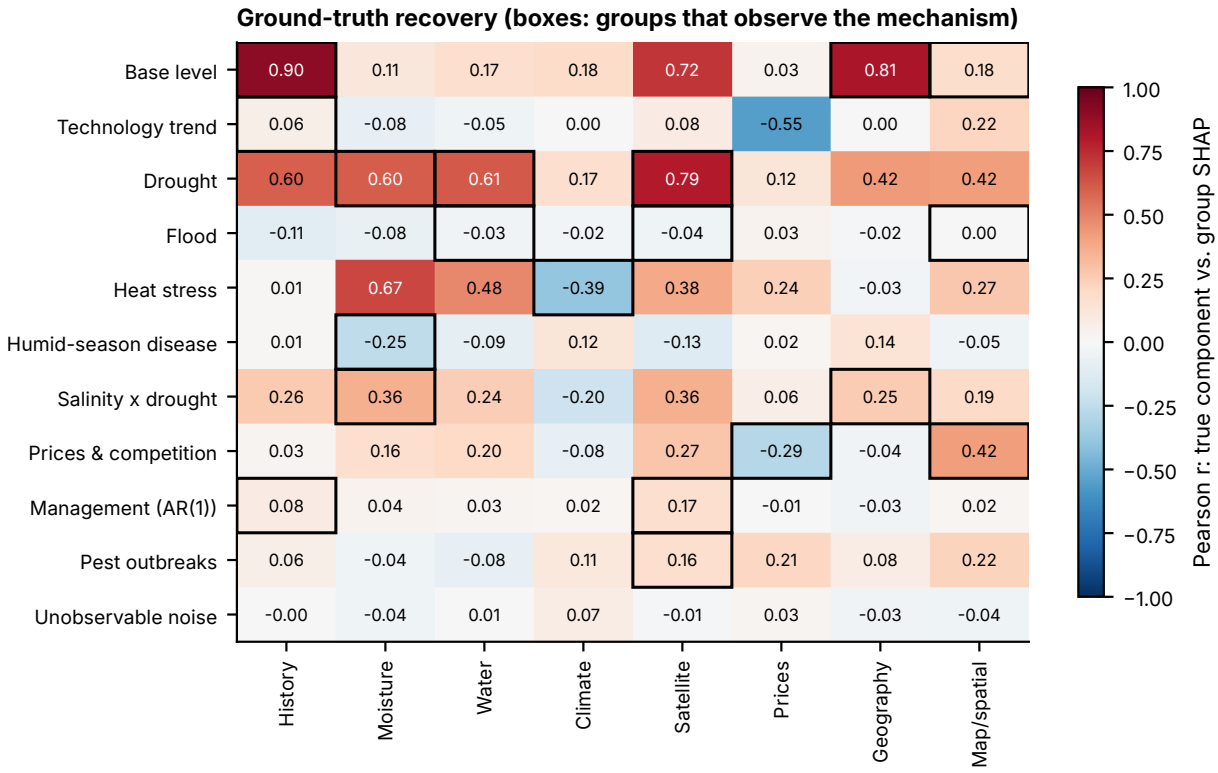


Figure 8: Correlation between each true yield component of the DGP and the summed SHAP values of each feature group (main model, 1,600 test forecasts). Boxes mark the groups that observe a mechanism; a faithful explanation would place large positive values in the boxes.

Table 5: Ground-truth recovery on the test forecasts. s.d.: standard deviation of the true component (kg rai^{-1}); r_{obs}: Pearson correlation between the true component and the summed SHAP values of the groups that observe it, for the main model and (post hoc) for LightGBM alone; last column (main model): the single group whose SHAP correlates most strongly (absolute value) with the component.

True component	s.d.	Observed by	r_{obs} main	r_{obs} LightGBM	Best single group (r)
Base level	44.8	Geography, Map/spatial, History	0.90	0.92	History (0.90)
Technology trend	5.0	History	0.06	0.06	Prices (-0.55)
Drought	39.0	Moisture, Water, Satellite	0.84	0.88	Satellite (0.79)
Flood	5.4	Water, Map/spatial, Satellite	-0.04	-0.05	History (-0.11)
Heat stress	9.5	Climate	-0.39	-0.34	Moisture (0.67)
Humid-season disease	5.1	Moisture	-0.25	0.00	Moisture (-0.25)
Salinity × drought	0.8	Geography, Moisture	0.41	0.33	Satellite (0.36)
Prices & competition	14.4	Prices, Map/spatial	-0.02	0.52	Map/spatial (0.42)
Management (AR(1))	17.4	History, Satellite	0.14	0.14	Satellite (0.17)
Pest outbreaks	10.3	Satellite	0.16	0.16	Map/spatial (0.22)
Unobservable noise	15.9	–	–	–	Climate (0.07)

5.6 Calibration of prediction intervals (RQ4)

Rolling split-conformal intervals of the main model were close to nominal: 80% intervals covered 78.3% and 90% intervals 86.6% of the test yields, with mean widths of 74 and 96 kg rai^{-1} (Fig. 11). Contrary to H6, coverage was *higher* in the shock seasons 2019, 2023 (85.2% and 93.2%) than in the other seasons (76.0% and 84.4%): the drought shocks were well predicted because their causes are observed. Under-coverage was instead concentrated in the season of the model failure (40% of 80% intervals in 2018). Calibration on past errors thus protects against average

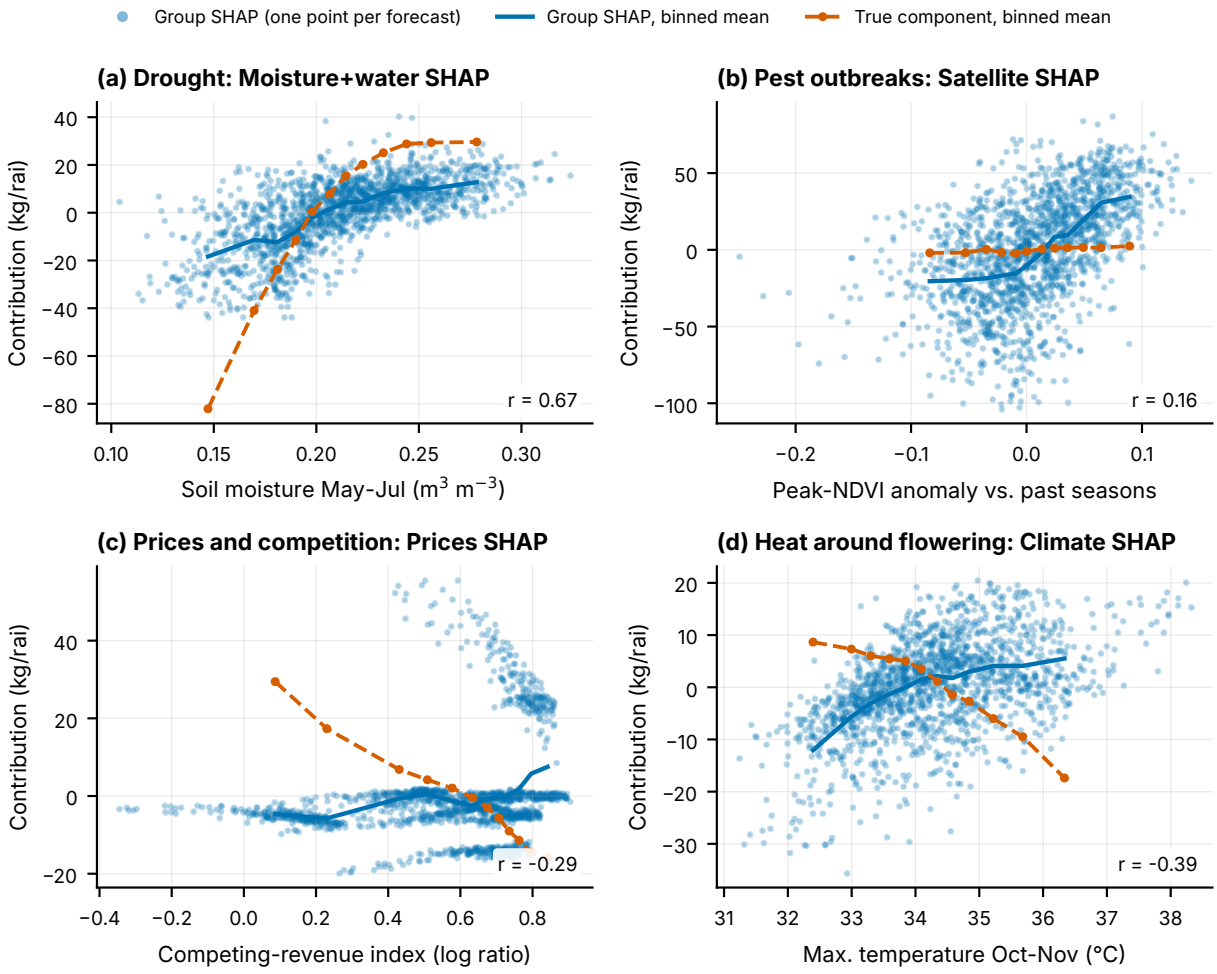


Figure 9: Group SHAP values of the main model (points and solid line: binned mean) and the true component (dashed line: binned mean) along the main observed driver of four mechanisms; both are centred. *r*: correlation between group SHAP and true component.

miscalibration but not against a new failure mode of the point forecaster.

5.7 Robustness across worlds and scenarios

Across ten independently simulated worlds (Table 6), LightGBM was remarkably stable (RMSE 26.5 $\pm$ 0.5, range 25.6–27.1), while climatology reached 49.4 $\pm$ 2.9. The ensemble had the lowest median RMSE (25.7) and beat LightGBM in 6 of 10 worlds, but its mean (31.8 $\pm$ 11.4) was inflated by 2 failures (RMSE more than 1.5 times that of LightGBM), and Ridge failed in 3 worlds (median 28.9, maximum 100.7). Without satellite features the ensemble's median RMSE rose to 33.6. In the scenario worlds LightGBM stayed between 26.3 and 28.6 kg rai^{-1}, even under strong non-linearity (28.5, against 87.2 for climatology), whereas Ridge degraded with non-linearity (157.5), cloud contamination (54.7) and map-layer error (39.3). In the world without pest outbreaks the three learners performed similarly (ensemble 25.0, Ridge 25.8, LightGBM 26.3).

5.8 Post-hoc diagnosis of the linear failures

The failures of the linear learner were not anticipated, so we diagnosed them after the confirmatory analysis; the results in this subsection are exploratory. RidgeCV selects its penalty by efficient leave-one-out cross-validation over *rows*,

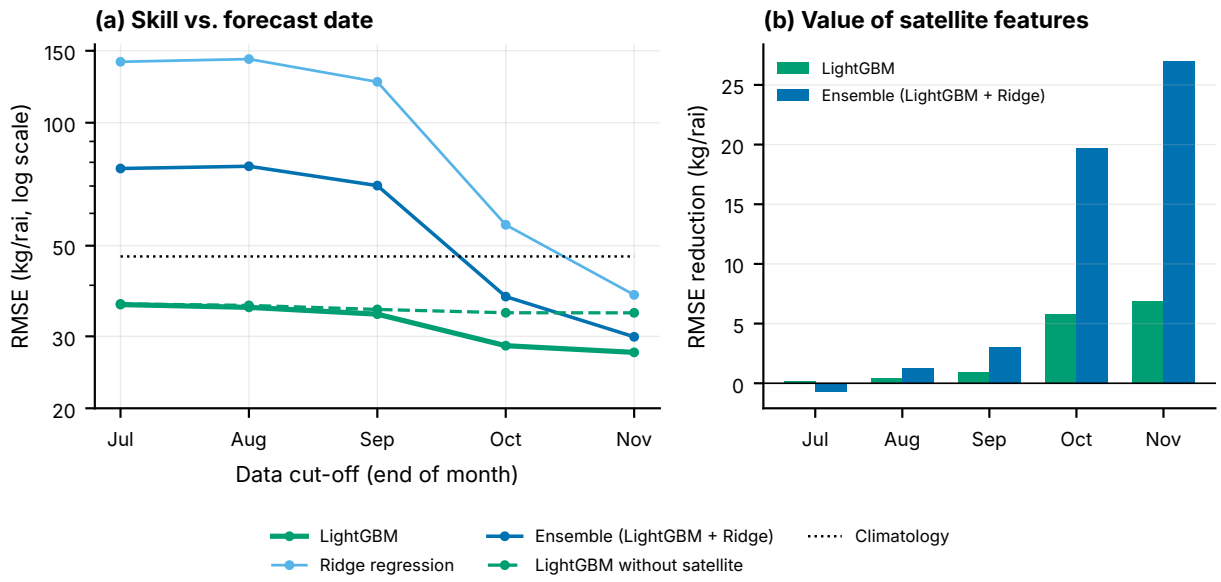


Figure 10: (a) Test RMSE as a function of the data cut-off month (log scale). (b) RMSE reduction contributed by the satellite group at each cut-off.

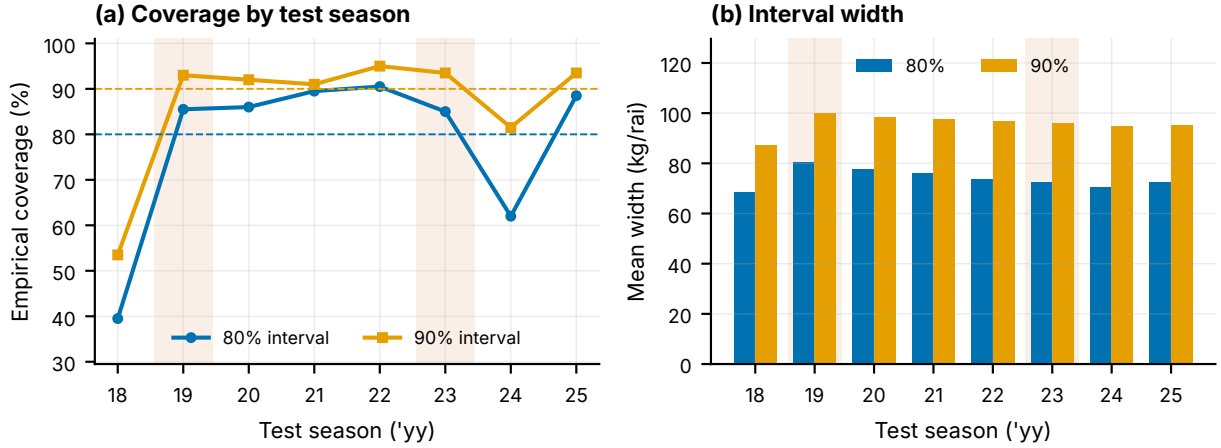


Figure 11: Rolling split-conformal prediction intervals of the main model: (a) empirical coverage by test season (dashed: nominal levels; shaded: shock seasons); (b) mean interval width.

which treats district-seasons as exchangeable. Because errors are strongly correlated within seasons, leaving out one district-season leaves its season effect visible through the other districts, and the selected penalty is small (in 2018: $\alpha = 0.11$; median over test folds 0.79). Selecting the penalty by leave-one-season-out (LOSO) cross-validation gave much stronger regularisation ($\alpha = 41.8$; median 62.1) and removed the failure (Table 7): the 2018 RMSE fell from 81.7 to 27.1 kg rai^{-1}, and the test RMSE of Ridge from 37.9 to 26.4. With the season-blocked penalty the ensemble would have reached 25.8 kg rai^{-1}, below LightGBM (27.4). The same mechanism explained the failures in the two worst replicate worlds and in the non-linear scenario. Tuning that ignores the season structure is thus itself a form of leakage—one that random outer cross-validation cannot detect.

Table 6: Robustness: test RMSE (kg rai^{-1}) in the main world, across ten independently simulated replicate worlds, and in four DGP scenarios (seed 42), with the configuration tuned once in the main world. "No sat.": ensemble without satellite features; "Random CV": the same ensemble scored with naive random 5-fold cross-validation.

World	Climatology	Ridge	LightGBM	Ensemble	No sat.	Random CV
Main world (seed 42)	47.1	37.9	27.4	29.9	56.9	23.8
Replicates, mean $\pm$ s.d. ($n = 10$)	49.4 $\pm$ 2.9	44.2 $\pm$ 28.6	26.5 $\pm$ 0.5	31.8 $\pm$ 11.4	38.5 $\pm$ 12.6	23.5 $\pm$ 0.4
Stronger non-linearity	87.2	157.5	28.5	83.9	166.8	26.2
3$\times$ map-layer error	47.1	39.3	27.6	30.5	55.4	23.8
1.7$\times$ cloud contamination	47.1	54.7	28.6	36.8	65.1	24.6
No satellite-only pests	46.8	25.8	26.3	25.0	43.6	23.6

Table 7: Post-hoc diagnosis (exploratory): test RMSE (kg rai^{-1}) of Ridge and of the LightGBM–Ridge ensemble when the Ridge penalty is selected by row-wise leave-one-out (as in the confirmatory analysis) or by leave-one-season-out (LOSO) cross-validation.

World	Ridge, row LOO	Ridge, LOSO	Ensemble, row LOO	Ensemble, LOSO	LightGBM
Main world (seed 42)	37.9	26.4	29.9	25.8	27.4
Replicate world 1002	91.7	27.7	50.5	25.7	26.0
Replicate world 1008	100.7	28.5	55.5	25.4	25.9
Non-linear scenario	157.5	32.0	83.9	27.8	28.5

Table 8: Summary of the evidence for the hypotheses.

Hypothesis	Evidence	Verdict
H1: multi-source learners beat the baselines	RMSE 27.4–29.9 vs. 47.1 (climatology); bootstrap CI of the gain excludes zero; replicated in all worlds; DM tests over 8 seasons not significant	Supported in effect size; tests underpowered
H2: flawed protocols are optimistic	random CV -10% (LightGBM) and concealed failures; current neighbour yields -5%; same-season prices and full-period anomalies negligible	Partly supported
H3: satellite unique, other groups substitute	satellite removal +25.0%; other groups between -2.8% and $+1.6\%$; unique value not limited to pests	Supported
H4: SHAP recovers strong, not weak or confounded mechanisms	base and drought $r \geq 0.84$; heat misattributed; trend absorbed by prices; price effect recovered only by the stable learner	Supported
H5: skill rises in season; satellite value after panicle initiation	LightGBM 35.9 $\rightarrow$ 27.4; satellite gain $\leq$0.9 until September, 6.8 in November	Supported
H6: intervals under-cover in shock seasons	78.3% overall; 85.2% in shock seasons; under-coverage in the failure season	Not supported

6 Discussion

6.1 What the benchmark teaches about multi-source yield forecasting

Three lessons stand out. The first is that *leakage-safe evaluation must reach every loop*. The most consequential failure in this study did not arise from the outer protocol, which was leakage-safe, but from an inner routine—the efficient leave-one-out selection of the Ridge penalty—that silently treated district-seasons as exchangeable. Within-season error correlation let season-level predictors such as prices and reservoir storage fit season effects, the penalty was set too small, and the model extrapolated badly when a new season brought new values of these predictors. Random outer cross-validation did not reveal the problem; it reported excellent and stable performance precisely because it never asked the model to forecast an unseen season. The practical rule is simple: every resampling step, inner and outer, should be blocked by season (and by space when forecasting unseen districts) (Roberts et al., 2017; Bergmeir and Benítez, 2012).

The second lesson is that *stability is a first-class model property*. Tree ensembles extrapolate conservatively, holding predictions constant beyond the training range, whereas linear models extrapolate along every predictor. With fewer than fifteen training seasons and several predictors that vary mainly between seasons, this asymmetry dominated: the tuned LightGBM varied by only 0.5 kg rai^{-1} across ten worlds, while Ridge failed in 3 of them. Ensembles inherit the failures of their members unless each member is individually robust; once Ridge was tuned by season, the ensemble became the most accurate learner (an exploratory result). The neural baseline did not beat the tree models, in line with benchmarks on tabular data (Grinsztajn et al., 2022; Shwartz-Ziv and Armon, 2022).

The third lesson is that *explanations need ground truth*. Grouped SHAP values reflected strong, well-observed mechanisms faithfully but misattributed correlated mechanisms (heat to moisture) and season-level effects (trend to prices), consistent with known difficulties of attributions under dependence (Aas et al., 2021; Hooker et al., 2021). Explanations also inherit model fragility: the ensemble attributed the largest share of its forecasts to prices, although removing prices improved it and its price attributions ran against the true response. In an empirical study such attributions could easily be read as evidence of economic effects. Grouped attributions of stable models, cross-checked against ablation and, where possible, against simulations with known truth, are safer; causal questions about prices require identification strategies from econometrics rather than predictive attributions (Wooldridge, 2010).

6.2 Implications for operational forecasting

If the benchmark's structure is representative of the region, a district-level system built from MODIS-class composites, ERA5-Land and CHIRPS weather, reservoir data and official statistics could issue informative forecasts from July and sharpen them substantially in October, when satellite information starts to add value, with 80% intervals about 74 $\mathrm{kg\,rai^{-1}}$ wide. Such a system would serve the agencies that compile yield statistics, monitor floods from space, manage reservoirs and administer crop insurance. The leakage audit implies that accuracy figures obtained with random cross-validation should not be used to justify such systems, and the failure analysis implies that model stability across seasons deserves as much attention as average accuracy.

6.3 Limitations

All conclusions are conditional on the DGP. Although its mechanisms are drawn from the agronomic and remote-sensing literature, the DGP is additive with hinge-type stresses, its parameters were set to plausible values rather than estimated from Thai data, and its satellite indices are generated from the visible part of yield; real indices also respond to weeds, mixed pixels, varietal and phenological differences, so the benchmark may overstate the value of satellite data. Only eight test seasons were available, which limits the power of season-level tests; replicate worlds address Monte-Carlo variability but not structural misspecification. The configuration was tuned once in the main world and reused elsewhere. SHAP values were computed in their interventional form; conditional variants may behave differently under dependence. The analysis is at district level and does not use field-level or radar data, and spatial cross-validation for districts without yield history was not evaluated. Finally, an exploratory version of this pipeline was run before the final design: although tuning and model selection used validation seasons only, the analyst was not blind to earlier test-period results, and the diagnosis of the linear failures is post hoc. Moreover, the post-hoc check was applied to worlds selected because they had failed; a prospective test of season-blocked tuning on fresh worlds is required before the remedy can be recommended on empirical grounds. Recovery was measured by correlation, which ignores the scale of attributions, and the neural baseline was not tuned.

6.4 Pathway to real data

The code ingests real inputs with the same schema (Table 2): yield statistics, CHIRPS and ERA5-Land weather, MODIS composites extracted with the provided Earth Engine script, farm-gate prices, reservoir storage and national soil, land-use and flood maps. Natural extensions are Sentinel-1 radar to observe the crop through monsoon clouds, Sentinel-2 for field-level models, spatially blocked validation for districts without history, conformalised quantile regression or adaptive conformal inference for heteroscedastic and shifting errors (Romano et al., 2019; Gibbs and Candès, 2021), and season-blocked inner tuning for every learner.

7 Conclusion

In a benchmark with known truth, multi-source machine learning reduced district-level rice yield forecast errors by 36–42% relative to climatology, with satellite data as the indispensable predictor group and with most of their value arriving from October. The same benchmark exposed three pitfalls that real-data studies cannot see: random cross-validation overstated skill and hid failures; leakage in an inner tuning loop caused catastrophic failures of a linear learner; and SHAP explanations were faithful only for strong, well-observed mechanisms and stable learners. We therefore recommend (i) blocking every resampling step by season, (ii) reporting skill relative to climatology and

on anomalies, (iii) preferring learners that are stable under season-level covariate shift, or tuning linear components by season—a remedy our exploratory analysis supports but that still needs prospective confirmation, (iv) treating attributions as descriptions of a model and validating them by simulation, and (v) reporting calibrated intervals together with their coverage by season. The benchmark and code are released so that these checks can precede the use of real data.

Data and code availability

All code, the simulated data, the saved results and the scripts that regenerate every number, table and figure of this paper are provided in the accompanying repository, `https://github.com/thanatchon36/isan-rice-yield-benchmark`, archived on Zenodo (doi:10.5281/zenodo.24036674); `make all` reproduces the analysis on a single CPU core. The simulated data and saved results are also published as a dataset, `https://huggingface.co/datasets/thanatchon36/isan-rice-yield-benchmark`.

AI-assistance statement

The study design, code, experiments and this manuscript were drafted with the assistance of a large language model (Claude, Anthropic) under the direction of the author, who reviewed the code and results and takes full responsibility for the content. Because the drafting model had no access to bibliographic databases, all references were compiled from memory and must be verified against the original sources before submission.

References

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

Bruno Basso and Lin Liu. Seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, 154:201–255, 2019.

I. Becker-Reshef, E. Vermote, M. Lindeman, and C. Justice. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment*, 114(6):1312–1323, 2010.

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

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Marshall Burke and David B. Lobell. Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences*, 114(9):2189–2194, 2017.

Yaping Cai, Kaiyu Guan, David Lobell, Andries B. Potgieter, Shaowen Wang, Jian Peng, Tianfang Xu, Senthold Asseng, Yongguang Zhang, Liangzhi You, and Bin Peng. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, 274:144–159, 2019.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232, 2001.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2, 2015.

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

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, 2022.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Dean P. Holzworth et al. APSIM – evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62:327–350, 2014.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31, 2021.

A. Huete, K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2):195–213, 2002.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition, 2021.

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. Random forests for global and regional crop yield predictions. *PLoS ONE*, 11(6), 2016.

David M. Johnson. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, 141:116–128, 2014.

J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A. Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman, and J. T. Ritchie. The DSSAT cropping system model. *European Journal of Agronomy*, 18(3–4):235–265, 2003.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 2012.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3): 1217–1241, 1989.

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

Corey Lesk, Pedram Rowhani, and Navin Ramankutty. Influence of extreme weather disasters on global crop production. *Nature*, 529(7584):84–87, 2016.

David B. Lobell and Marshall B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010.

David B. Lobell, Wolfram Schlenker, and Justin Costa-Roberts. Climate trends and global crop production since 1980. *Science*, 333(6042):616–620, 2011.

David B. Lobell, David Thau, Christopher Seifert, Eric Engle, and Bertis Little. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164:324–333, 2015.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Christoph Molnar. *Interpretable Machine Learning*. Independently published, 2nd edition, 2022.

Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

Joaquín Muñoz-Sabater et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021.

Natural Earth. Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales. https://www.naturalearthdata.com, 2026. Public domain; GeoJSON files accessed 7 November 2026.

Marc Nerlove. Estimates of the elasticities of supply of selected agricultural commodities. *Journal of Farm Economics*, 38(2):496–509, 1956.

Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Deepak K. Ray, James S. Gerber, Graham K. MacDonald, and Paul C. West. Climate variation explains a third of global crop yield variability. *Nature Communications*, 6, 2015.

Angela J. Rigden, Nathaniel D. Mueller, N. Michele Holbrook, Neel Pillai, and Peter Huybers. Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting US maize yields. *Nature Food*, 1(2):127–133, 2020.

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. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

C. F. Ropelewski and M. S. Halpert. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, 115(8):1606–1626, 1987.

Wolfram Schlenker and Michael J. Roberts. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009.

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

Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

Compton J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Zhengming Wan. New refinements and validation of the MODIS land-surface temperature/emissivity products. *Remote Sensing of Environment*, 112(1):59–74, 2008.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edition, 2010.

Xiangming Xiao, Stephen Boles, Jiyuan Liu, Dafang Zhuang, Steve Frolking, Changsheng Li, William Salas, and Berrien Moore. Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. *Remote Sensing of Environment*, 95(4):480–492, 2005.

Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep Gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

A Benchmark data-generating process

Random fields are unit-variance, kernel-smoothed white noise, $g_\ell = \mathrm{std}(K_\ell \zeta)$ with $\zeta \sim \mathcal{N}(0, I)$ and $K_\ell(i,j) = \exp\{-d_{ij}^2/(2\ell^2)\}$, where d_{ij} is the distance between district centroids in km; ν denotes independent standard normal draws.

Rainfall. $\mathrm{rain}_{itm} = N_i \varphi_m \exp(z_{itm} - 0.065)$ with $z_{itm} = -0.2\, e_t c_m + 0.18 A_{it} + 0.30 M_{itm} + 0.08 \nu_{itm}$, where N_i is the rainfall normal, $\varphi_m = (0.11, 0.13, 0.14, 0.18, 0.20, 0.09, 0.02)$ for May–November, $c_m = 1.4$ in June–July and 1 otherwise, and A_{it} and M_{itm} are fields with $\ell = 150$ and 80 km.

Temperature and humidity. With $\tilde{z} = \mathrm{clip}(z/0.36, -3, 3)$ and $\theta_{it} = 0.30 e_t + 0.25 B_{it}$ (B a 150-km field): $T_{itm} = \bar{T}_m - 0.0065(h_i - 170) + 0.035(t - 2008) + \theta_{it} - 0.35\tilde{z}_{itm} + 0.15\nu$; $T^{\max}_{itm} = T_{itm} + 6 + 0.8\max(0, -\tilde{z}_{itm}) + 0.3\nu$; and $\mathrm{RH}_{itm} = \mathrm{clip}(\overline{\mathrm{RH}}_m + 4\tilde{z}_{itm} - 1.5\theta_{it} + 1.5\nu,\, 40,\, 98)$, where h_i is elevation and $\bar{T}_m$, $\overline{\mathrm{RH}}_m$ are regional monthly climatologies.

Soil moisture and reservoirs. $s_{itm} = 0.45\, s_{it,m-1} + 0.55\{0.08 + 0.42(1 - e^{-\mathrm{rain}_{itm}/200})(1.30 - \mathrm{sand}_i/100)\}$; the observed series adds $\mathcal{N}(0, 0.025^2)$ error. Storage at sowing is $\mathrm{clip}(58 - 14 e_{t-1} + \delta_b + 7\nu,\, 12,\, 95)\%$ with $\delta_b = -8$, 0 and 6 for the Mekong, Mun and Chi basins, and the irrigation supply is $S_{it} = I_i \times \mathrm{storage}_{b(i)t}/100$.

Prices. $\ln P_t = \mu + 0.6(\ln P_{t-1} - \mu) + \Sigma^{1/2}\nu_t$ plus occasional known shocks, with $\mu = \ln(10.5, 2.2, 1.0, 8.0, 14.0)$ THB kg^{-1} for rice, cassava, sugarcane, maize and urea, standard deviations $(0.11, 0.17, 0.11, 0.13, 0.13)$, and correlations of 0.35 among crops and 0.2 with urea. Local prices are $P_{c,t-1} - \kappa_c d_{ic}$ with $\kappa = (0.010, 0.006, 0.004, 0.010)$ THB kg^{-1} km^{-1} and the distance to a market or, for cane, to a sugar mill. The competing-revenue index is

$$c_{it} = \ln \frac{p_{\mathrm{cas}} \bar{Y}_{\mathrm{cas}} + w_i\, p_{\mathrm{cane}} \bar{Y}_{\mathrm{cane}} + p_{\mathrm{mz}} \bar{Y}_{\mathrm{mz}}}{(2 + w_i)\, p_{\mathrm{rice}} \bar{Y}_{\mathrm{rice}}},$$

with $w_i = e^{-d_i^{\mathrm{mill}}/40}$ and typical yields $\bar{Y} = (400, 3500, 11000, 700)$ kg rai^{-1} for rice, cassava, sugarcane and maize; the input index is $\ln(p_{\mathrm{rice}}/P_{\mathrm{urea}})$. Both indices are standardised before entering Table 1. After each season the rice price is multiplied by $\exp(-\kappa \bar{a}_t/100)$, where $\bar{a}_t$ is the regional mean yield deviation from the base level and trend.

Satellite indices. With visible yield v_{it} (all components except noise), $\mathrm{NDVI}^{\mathrm{rice}}_{itm} = 0.20 + (N^{\mathrm{pk}}_{it} - 0.20)\exp\{-((m - m^{\mathrm{pk}}_{it})/1.7)^2\}$, $N^{\mathrm{pk}}_{it} = 0.42 + 0.38/\{1 + \exp(-(v_{it} - 380)/85)\}$ and $m^{\mathrm{pk}}_{it} = 9.4 + 0.8\,\mathrm{clip}((0.25 - \bar{s}^{MJ}_{it})/0.05, 0, 1.5)$; flooding reduces September–October values by up to 25%, and observed NDVI is $0.85\,\mathrm{NDVI}^{\mathrm{rice}} + 0.15 \times 0.62 + \mathcal{N}(0, 0.025^2)$. EVI equals $0.12 + (\mathrm{EVI}^{\mathrm{pk}} - 0.12)$ times the same seasonal shape plus $\mathcal{N}(0, 0.02^2)$, with $\mathrm{EVI}^{\mathrm{pk}} = 0.22 + 0.00075(v - 250)$; LSWI is $-0.10 + 0.9(s - 0.2)$ plus transplanting and flood terms and $\mathcal{N}(0, 0.03^2)$; LST is $T^{\max} + 1.5 + 25(0.28 - s) - 6(\mathrm{NDVI} - 0.5) + \mathcal{N}(0, 0.8^2)$. With monthly cloud probabilities $p_m = (0.15, 0.35, 0.45, 0.50, 0.45, 0.25, 0.10)$, a composite is lost with probability $0.25p_m$ and contaminated with probability $0.75p_m$ (NDVI reduced by $U(0.05, 0.18)$, EVI by $U(0.03, 0.10)$, LST by $U(2, 6)$ °C, LSWI increased by $U(0, 0.08)$).

B Additional results

This appendix documents the tuning of LightGBM and gives the detailed numbers behind Fig. 5, Fig. 7 and Table 6. Table B.1 lists the random-search grid, the configuration selected on the validation seasons and the validation RMSE of the three candidate main models. Table B.2 gives the test RMSE of every learner in every season: the failures of the linear learners (Ridge in 2018 and the panel fixed-effects model in 2020) stand out against otherwise competitive errors, and outside these seasons the ensemble is usually the most accurate learner. Table B.3 reports the drop-one-group ablation for all three learners: for LightGBM only the satellite group matters, whereas for Ridge and the ensemble the removal of most groups *lowers* the error—a symptom of the extrapolation failures diagnosed in Section 5.8. Table B.4 lists every replicate and scenario world; the failures of Ridge, and of the ensemble that contains it, are confined to a few worlds, whereas LightGBM's error varies little.

Table B.1: LightGBM random search (15 configurations × 2 target formulations, scored by rolling origin on 2015–2017) and the selected configuration (target: level). Validation RMSE of the candidate main models (kg rai^{-1}): Ensemble (LightGBM + Ridge) 26.2 ; Ridge regression 27.4 ; LightGBM 28.1.

Hyper-parameter	Search grid	Selected
`num_leaves`	7, 15, 31, 63	63
`learning_rate`	0.01, 0.02, 0.03, 0.05, 0.08	0.05
`n_estimators`	200, 400, 700, 1000	400
`min_child_samples`	10, 20, 40, 80	80
`subsample`	0.6, 0.8, 1.0	0.8
`colsample_bytree`	0.3, 0.5, 0.7, 0.9	0.5
`reg_lambda`	0.0, 1.0, 5.0, 20.0	1
`max_depth`	-1, 4, 6, 8	-1

Table B.2: Test RMSE (kg rai^{-1}) of every learner in every test season (bold: lowest error of the season) and over all 1,600 forecasts.

Model	2018	2019	2020	2021	2022	2023	2024	2025	All
LightGBM	28.7	27.0	27.0	26.4	26.1	27.1	31.9	24.4	27.4
Random forest	27.4	25.7	28.9	28.6	26.9	29.2	32.3	25.9	28.2
Ensemble (LightGBM + Ridge)	50.9	**24.8**	**26.3**	**25.6**	**24.2**	24.3	29.5	**23.7**	29.9
LightGBM (alternative target)	28.0	34.0	28.5	29.0	28.0	33.7	36.7	30.2	31.2
Multilayer perceptron	35.8	30.3	35.9	29.7	33.7	30.3	33.0	27.8	32.2
Panel fixed effects	**27.1**	27.9	65.4	27.1	27.3	29.0	34.7	24.0	35.2
Ridge regression	81.7	25.3	27.2	28.2	25.2	**24.3**	**28.5**	24.8	37.9
Climatology (3-season mean)	29.9	49.3	43.4	54.5	42.4	62.6	47.2	39.8	47.1
Persistence	44.8	51.1	64.0	58.0	37.2	71.1	84.1	50.1	59.2

Table B.3: Drop-one-group ablation: change in test RMSE (%) when a feature group is removed and the learner re-trained in every fold; brackets: 95% season-block bootstrap interval. Positive values mean the group is useful. The first row gives the RMSE (kg rai^{-1}) with all features.

Removed group	LightGBM	Ridge	Ensemble
(none: RMSE with all features)	27.4	37.9	29.9
Satellite (NDVI, EVI, LSWI, LST)	+25.0 [+13.0, +41.3]	+153.3 [+12.9, +279.8]	+90.3 [+14.1, +170.6]
Map layers & spatial lags	+1.6 [-0.5, +3.7]	-22.2 [-67.5, +12.9]	-9.5 [-34.0, +6.9]
Geography (terrain, soil)	+1.1 [-0.7, +2.7]	+30.0 [-0.4, +60.0]	+15.1 [-0.5, +34.3]
Prices (own & competing crops)	+0.4 [-1.3, +2.2]	-31.4 [-74.4, +0.1]	-13.9 [-37.9, +1.3]
Humidity & soil moisture	-0.6 [-2.9, +1.8]	-30.4 [-74.4, +1.9]	-14.1 [-39.3, +1.6]
Climate (temperature, GDD)	-0.7 [-2.3, +1.0]	-28.5 [-70.1, +2.3]	-13.2 [-36.0, +1.4]
Water (rain, reservoir, irrigation)	-0.9 [-3.1, +1.0]	-8.3 [-20.3, +3.8]	-4.7 [-13.1, +2.3]
Previous yields	-2.8 [-4.8, -0.7]	-16.3 [-42.4, +6.5]	-8.2 [-22.8, +1.9]

Table B.4: Test RMSE (kg rai^{-1}) in every replicate world (seeds 1000–1009) and scenario world (seed 42), using the configuration tuned once in the main world. †: RMSE more than 1.5 times that of LightGBM in the same world. No sat.: ensemble without satellite features; Random CV: the same ensemble scored by random 5-fold cross-validation.

World	Climatology	Ridge	LightGBM	Ensemble	No sat.	Random CV
Replicate world 1000	46.4	26.2	26.2	24.9	31.4	23.4
Replicate world 1001	48.1	26.0	25.6	24.4	31.0	23.4
Replicate world 1002	52.4	91.7†	26.4	50.3†	70.2	23.3
Replicate world 1003	46.8	27.3	27.1	25.9	30.1	24.4
Replicate world 1004	49.2	29.7	26.3	25.6	38.0	23.6
Replicate world 1005	54.2	34.3	27.1	28.4	39.2	23.5
Replicate world 1006	47.9	51.8†	27.0	33.9	48.4	23.7
Replicate world 1007	53.4	28.1	26.8	24.9	29.3	23.1
Replicate world 1008	48.1	100.7†	26.2	55.2†	35.6	23.1
Replicate world 1009	47.0	26.1	26.1	25.1	31.6	23.6
Stronger non-linearity	87.2	157.5†	28.5	83.9†	166.8	26.2
3× map-layer error	47.1	39.3	27.6	30.5	55.4	23.8
1.7× cloud contamination	47.1	54.7†	28.6	36.8	65.1	24.6
No satellite-only pests	46.8	25.8	26.3	25.0	43.6	23.6