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.

An explainable digital twin for balancing food production, resource use and post-harvest waste: a groundnut case study from Tamil Nadu

Arshed Ayoob
Vellore Institute of Technology
Yoganand Selvaraj
s.yoganand@vit.ac.in

Vellore Institute of Technology
ArunPandian J
Vellore Institute of Technology

Research Article

Keywords: Digital Twin, Explainable AI, SHAP, NSGA-II, Groundnut, Tamil Nadu, Multi-objective optimization, Post-harvest waste

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11589751/v1

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

Additional Declarations: No competing interests reported.

An explainable digital twin for balancing food production, resource use and post-harvest waste: a groundnut case study from Tamil Nadu

Arshed Ayoob P[1], Yoganand S[2,*], Arun Pandian J[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore 632014, Tamil Nadu, India

[2]Centre for Artificial Intelligence and Research (CFAIR), Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India

Emails: arshedayoob.p2025@vitstudent.ac.in, s.yoganand@vit.ac.in, arunpandian.j@vit.ac.in

[*]Corresponding author: Yoganand S, s.yoganand@vit.ac.in

Abstract

Food planners trade marketable output against water, fertilizer and post-harvest losses, and yield models predict output without pricing those trades. We built a computational digital twin of groundnut production in 36 Tamil Nadu districts. A Ridge regression surrogate predicts district production from cultivated area and previous production. SHapley Additive exPlanations (SHAP) attribute its predictions to the inputs, and the non-dominated sorting genetic algorithm II (NSGA-II) optimizes four objectives: marketable food, water use, a fertilizer-resource index and post-harvest waste. The district records sum to the official 2022–23 state totals within 0.001%. In repeated cross-validation the surrogate reached a coefficient of determination of 0.920±0.058 and a mean absolute error of 3,923 t, and adding weather variables lowered the score to 0.916. Previous production outweighed cultivated area, with mean absolute SHAP values of 15,321 t and 6,471 t. Dropping the waste objective cut the normalized hypervolume from 0.39177 to 0.32708. The balanced Pareto solution reduced modelled water use by 21.23% and waste by 60.60%, and it cost 20.81% of marketable food. Harvested output fell by 21.25%, so the water saving comes from producing less, and the waste saving comes from an assumed halving of the loss rate. All reductions are simulated and no field measurement supports them.

Keywords: Digital Twin; Explainable AI; SHAP; NSGA-II; Groundnut; Tamil Nadu; Multi-objective optimization; Post-harvest waste

1 Introduction

Feeding more people from the same land means raising usable output while holding down water use, fertilizer use and post-harvest loss. These goals pull against each other: more area and more input raise production, and less input lowers it. Machine learning predicts crop yield from temperature, rainfall and soil variables [1], and the same reviews that document this progress list scarce data, poor transfer between regions and low interpretability as open problems [1, 2].

A digital twin couples a model to an observed state so that planners can test scenarios before acting. Agricultural twins have been built from crop, soil, weather and sensor data [3–6]. Purcell and Neubauer judge what-if simulation and decision optimization to be less developed than monitoring [5], and Purcell et al. ask for explicit checks on data quality and model fidelity [7].

Groundnut in Tamil Nadu has district-level records and an existing yield literature. Krithika et al. modelled district groundnut yield from irrigation, rainfall, area and production and found that the choice of features changed accuracy [8]. Their model ends at the yield prediction. It does not test scenarios for water, fertilizer or post-harvest loss, and it does not search across competing objectives.

This paper builds a twin that takes both steps for 36 districts. The twin is a computational model refreshed from scenario inputs, without a live sensor feed, and it adds no new optimization algorithm. A Ridge surrogate predicts production, SHapley Additive exPlanations (SHAP) attribute each prediction to its inputs, and the non-dominated sorting genetic algorithm II (NSGA-II) searches for trade-offs. The evaluation asks five questions. RQ1: which model predicts district production most stably? RQ2: does district weather improve it? RQ3: which inputs drive its predictions? RQ4: which trade-offs among marketable food, water, fertilizer-resource use and post-harvest waste does the optimizer expose? RQ5: does keeping all four objectives change the Pareto set compared with dropping one?

The paper contributes an architecture, named AgriFoodTwin-XAI after its explainable artificial intelligence layer, that links data, surrogate, scenario model and optimizer. It documents a 36-district dataset whose totals match the official 2022–23 figures within 0.001%, and it states the four-objective problem with explicit response equations and bounds. It also separates the balanced scenario into the part caused by lower output and the part caused by assumed response functions.

The remainder of the article is organized as follows. Section 2 reviews related work. Section 3 describes the framework and the response models, Section 4 the data and experimental design, and Section 5 the results. Section 6 discusses the results, Section 7 states the limitations, and Section 8 concludes.

2 Related work

Van Klompenburg et al. reviewed crop yield prediction and found temperature, rainfall and soil type to be the most used predictors, with neural networks prominent among the models [1]. Benos et al. report the same for precision agriculture and add data heterogeneity and weak generalization as persistent problems [2]. Satellite data widen the feature set, but accuracy depends on temporal coverage, spatial resolution and ground reference data [9]. A deep network would have 36 annual records to train on. On medium-sized tabular benchmarks, tree ensembles outperform deep networks [10], and a regularized linear model [11] is the usual baseline when samples are fewer still.

Verdouw et al. give a conceptual framework for twins in smart farming [3], and Pylianidis et al. report that agricultural applications are still at an early stage [4]. Purcell and Neubauer define an agricultural twin by state sensing, entity understanding, modelling and, possibly, bidirectional data streams [5]. Peladarinos et al. describe twins that combine soil, crop, weather, sensor and image data for irrigation and fertilization decisions [6]. Purcell et al. add the sustainability risks and data-quality demands that such twins carry [7]. Our twin is computational and refreshed from scenario inputs, and it has no continuous link to a physical farm.

SHAP assigns each input an additive share of a prediction [12], and an exact algorithm for tree models makes the computation fast [13]. Rudin argues that high-stakes decisions call for interpretable models instead of explanations added to black boxes [14]. We take that advice: a Ridge model can be read from its coefficients, and SHAP expresses them in tonnes.

NSGA-II approximates a Pareto set through non-dominated sorting and elitism, which avoids a single weighted score [15]. Chen et al. applied it to drip irrigation of citrus for yield, fruit quality and water-use efficiency [16], and Liu and Yang to irrigation scheduling with a spatially distributed crop model [17]. Wang et al. used machine-learning surrogates to find water and nitrogen rates for winter wheat [18]. The FAO yield-response factor K_y represents water stress [19, 20], and AquaCrop builds on the same relationship [21].

Post-harvest loss is the fourth objective. Assessments of Indian crops and of global supply chains find losses at several stages, with different rates by commodity [22–24]. Treating loss as a scenario variable avoids the assumption that every harvested tonne becomes food. Each component of the twin is established, and the contribution is their integration for one crop. Table 1 compares that integration with the cited literature.

Table 1 Positioning of AgriFoodTwin-XAI against the cited literature

Dimension	Typical prior work	AgriFoodTwin-XAI
Primary emphasis	Prediction, monitoring or conceptual twin design [1, 3, 4]	Production-state decision support
Digital Twin role	Virtual representation and simulation [5, 6]	Scenario-refreshable computational state
Explainability	Usually reported separately from optimization	SHAP attached to the surrogate used by the twin
Decision mechanism	Varies; NSGA-II in irrigation studies [16, 17]	NSGA-II Pareto optimization with four objectives
Objectives	One or a few resource variables	Food, water, fertilizer resource and waste
Empirical unit	Varies	36 Tamil Nadu groundnut districts

3 The AgriFoodTwin-XAI framework

The framework has five layers. The data layer holds district area, previous production and weather. The surrogate layer predicts production from them, and the XAI layer attributes each prediction to its inputs. The twin core combines observed

and predicted quantities into scenario states, and the optimization layer searches the space of interventions for non-dominated decisions. Fig. 1 shows the layers and three information flows. Observed data flow into the state, the surrogate feeds SHAP, and decisions pass through the twin to produce objective values. Only the last flow is optimized. The surrogate is fixed during a run.

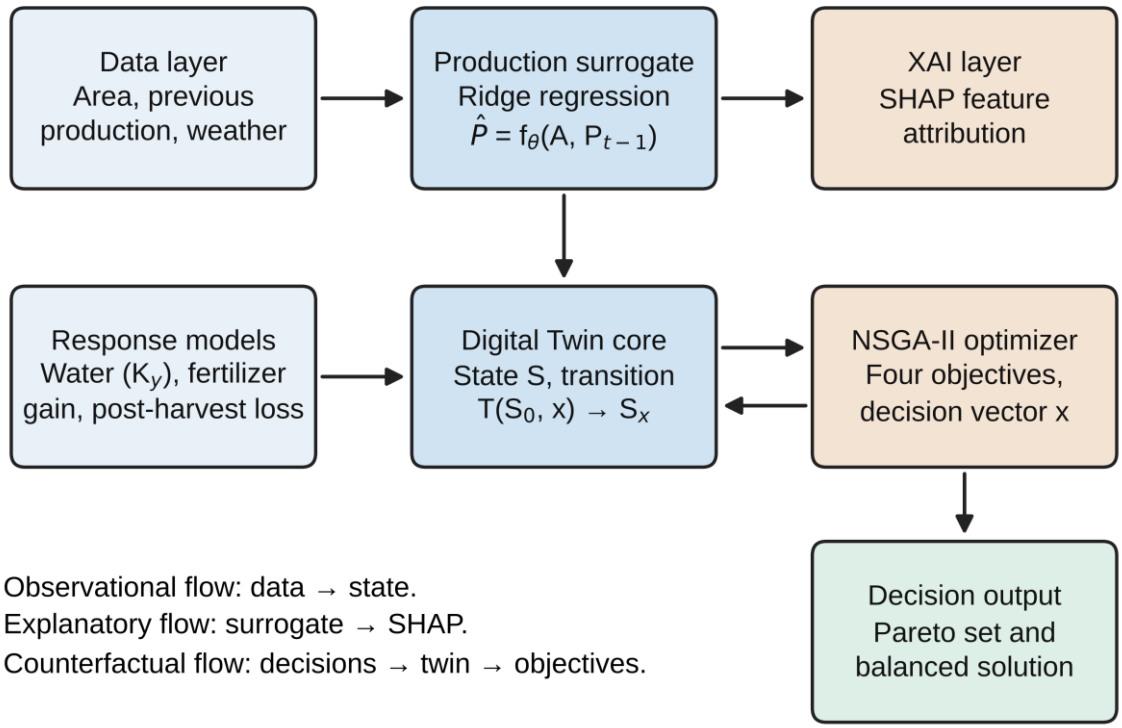


Fig. 1 AgriFoodTwin-XAI architecture and its three information flows

The twin holds a state for every district. For district d the state is S_d = {A_d, $P_{d,t-1}$, W_d, F_d, L_d}, where A is cultivated area, P_{t-1} is production in the previous period, W is the normalized water scenario, F is the normalized fertilizer-resource scenario and L is the post-harvest loss state. The surrogate predicts production as

$$\hat{P}_d = f_\theta(A_d, P_{d,t-1}, X_{\text{weather,d}}), \quad (1)$$

where $X_{weather}$ holds precipitation, temperature, relative humidity and solar-radiation summaries and enters only the weather configuration. The surrogate is trained on production alone. Separate response functions turn its output into counterfactual scenarios, because the district data contain no measured effect of any intervention.

The optimizer acts on a decision vector x = [x_A, x_F, x_W, x_L] with bounds $l_j \le x_j \le u_j$. It maximizes marketable production M(x) and minimizes water consumption W(x), the fertilizer-resource index F(x) and post-harvest waste L(x). In minimization form,

$$\min F(x) = [-M(x), W(x), F(x), L(x)], \quad x_j \in [l_j, u_j]. \quad (2)$$

Harvested production P(x) splits into marketable food and waste, M(x) = P(x) − L(x), with L(x) = ℓ(x)P(x), where ℓ(x) is the effective post-harvest loss rate. The twin applies the transition T(S_0, x) → S_x. The production component estimates output, the water and fertilizer modules transform resource demand, and the post-harvest module converts harvested production into food and waste. NSGA-II returns several non-dominated alternatives and needs no prior weights among the objectives [15]. A balanced solution is picked after the Pareto set exists, by the rule in Section 4.

The response models turn the decision vector into resource use and output. Four dimensionless factors form the vector: an area factor x_A, a water factor x_W, a fertilizer factor x_F and a post-harvest control level x_L. Table 2 lists their bounds. Water stress follows the crop-water relationship of the Food and Agriculture Organization (FAO)

$$\frac{Y_a}{Y_m} = 1 - K_y\left(1 - \frac{ET_a}{ET_m}\right), \tag{3}$$

where Y_a and Y_m are actual and maximum yield, ET_a and ET_m are actual and maximum evapotranspiration, and K_y is the crop sensitivity to water deficit [19, 20]. In the twin, the water factor plays the role of ET_a/ET_m, and the yield response is

$$R_w = \mathrm{clip}[1 - K_y(1 - x_W), 0.30, 1.0], \tag{4}$$

The main scenario uses K_y = 0.63 for groundnut, and the sensitivity analysis spans 0.50 to 0.90. The district data give no estimate of K_y. They also hold no fertilizer application data, so the fertilizer effect is a bounded gain factor with fixed parameters:

$$R_f = 1 + 0.088\,\frac{\tanh(z)}{\tanh(1)}, \quad z = \mathrm{clip}[\frac{x_F - 1}{0.10}, -2, 1]. \tag{5}$$

Production under decision x multiplies both responses onto the surrogate output, evaluated after scaling cultivated area by x_A:

$$P(x) = \hat{P}(x_A A, P_{t-1})\, R_w R_f. \tag{6}$$

The post-harvest module lowers the baseline loss rate ℓ_0 in proportion to the control level, by at most 50%:

$$\ell(x) = \ell_0(1 - 0.50\, x_L), \tag{7}$$

The baseline ℓ_0 is 1.11%, the producer-level groundnut post-harvest loss estimate for Tamil Nadu used in this study. The increased-loss case uses 6.03%, a national benchmark for overall harvest and post-harvest loss from the Indian Council of Agricultural Research–Central Institute of Agricultural Engineering (ICAR-CIAE). It applies to India as a whole and serves as an uncertainty reference. Water consumption and the fertilizer-resource index are

$$W(x) = A_{\mathrm{base}} x_A\, d\, x_W \times 10, \quad F(x) = x_F, \tag{8}$$

where A_{base} is the baseline cultivated area in hectares, d = 740 mm is the reference water depth, and 1 mm over 1 ha equals 10 m^3. The baseline area of 351,083 ha gives 2.598×10^9 m^3 in Eq. 8. All response functions are literature-informed scenario

functions. No measured irrigation, fertilizer or post-harvest data from the 36 districts calibrate them.

Table 2 Decision variables and bounds

Decision variable	Symbol	Range	Role in the twin
Area factor	x_A	0.90–1.05	Scales cultivated area before the surrogate prediction
Water factor	x_W	0.70–1.05	Water-deficit response in Eq. 4 and water use in Eq. 8
Fertilizer factor	x_F	0.80–1.10	Bounded gain in Eq. 5 and fertilizer-resource index
Post-harvest control	x_L	0–1	Loss-rate reduction in Eq. 7

4 Data and experimental design

Section 4 describes the dataset, the surrogate models with their validation protocol, and the settings of the explanation and optimization steps.

The dataset has 36 district-level groundnut observations for Tamil Nadu: cultivated area, current production and previous production. The target is production in 2022–23, and previous production is the output of the year before. We reconstructed the district values for this study and compared their totals with the official state figures before modelling. Table 3 gives the comparison. Area totals 351,083 ha against 351,084 ha officially, and production totals 912,148 t against 912,150 t, relative differences of 0.000285% and 0.000219%. The official figures come from the Statistical Handbook of Tamil Nadu 2022–23 [25]. Matching totals show that the districts add up to the published state figures. They say nothing about the accuracy of any single district.

Table 3 Reconstructed district totals compared with the official reference totals

Quantity	Reconstructed	Official reference	Relative difference
Area	351,083 ha	351,084 ha	0.000285%
Production	912,148 t	912,150 t	0.000219%

We compared two input sets. The historical set uses cultivated area and previous production. The weather set adds precipitation, minimum, maximum and mean temperature, relative humidity and solar radiation from NASA Prediction of Worldwide Energy Resources, abbreviated POWER [26]. We downloaded daily values for 1 June 2022 to 31 May 2023, summed precipitation over the period and averaged the other variables. Each district got one latitude–longitude point from the Nominatim geocoder.

Five regressors were tested: Ridge [11], Random Forest, Extra Trees, Gradient Boosting and HistGradientBoosting, all from scikit-learn [27]. Ridge is the linear baseline, and the four tree ensembles test whether non-linear structure helps. No hyperparameter search was run. Ridge uses standardization and $\alpha = 1.0$ in a pipeline, so scaling is fitted on each training fold only. Random Forest and Extra

Trees use 500 trees. Gradient Boosting uses 250 estimators, a learning rate of 0.03 and a maximum depth of 2. HistGradientBoosting uses 250 iterations, a learning rate of 0.05 and at most 6 leaf nodes. All models use random_state 42 and scikit-learn defaults elsewhere.

A six-fold KFold split first produced out-of-fold predictions for diagnostics. Model comparison used ten repetitions of six-fold RepeatedKFold, which gives 60 held-out evaluations per configuration. We scored each evaluation by the coefficient of determination R^2, the mean absolute error MAE and the root mean square error RMSE, and we report the mean and standard deviation SD over the folds. A held-out fold holds six districts, so a single-fold R^2 is noisy, and the 60-fold average is the figure we report. RepeatedKFold replaces cross_val_predict for the comparison because cross_val_predict needs every observation to fall in exactly one test fold.

SHAP values were computed for the selected Ridge surrogate, and global importance is the mean absolute SHAP value of each feature. NSGA-II ran on the four-objective twin with a population of 80, 100 generations and seed 42. The initial population was drawn uniformly within the bounds in Table 2. Parents were chosen by binary tournament on non-dominated rank and crowding distance. Offspring came from arithmetic crossover with complementary convex combinations of the parents, and each variable was mutated with probability 0.20 by Gaussian noise with a standard deviation of 0.04 times its range. Variables were clipped to their bounds, and survivors were taken from the combined parent and offspring population by rank and crowding distance.

Three ablations removed one objective each: waste, water or fertilizer. Every ablated run was scored on all four objectives, including the removed one, so the hypervolumes can be compared. We estimated hypervolume [28] by Monte Carlo sampling of 200,000 points with seed 42. The objectives were min–max normalized over the pooled candidate solutions, and the reference point was 1.05 in every normalized objective. The baseline is the observed state with a loss rate of 1.11%. After each Pareto set was generated, we normalized its four objectives and selected the solution with the lowest score, defined as the sum of the squared normalized deviations of production from its best value and of water, fertilizer and waste from zero. Sensitivity runs varied K_y and the loss rate. A bootstrap over the out-of-fold predictions gave interval estimates for the error metrics.

5 Results

Section 5.1 selects and explains the production surrogate using prediction error on held-out districts. Sections 5.2 and 5.3 report outputs of the twin: the Pareto solutions, the ablations and the balanced scenario, followed by sensitivity to the assumed parameters and the sampling uncertainty of the error metrics.

5.1 Surrogate performance and explanation

Ridge on the historical inputs predicted best and varied least across folds. Its R^2 was 0.920±0.058, its MAE 3,923 t and its RMSE 5,298 t. Gradient Boosting and Extra

Trees reached R^2 of 0.838 and 0.834 with standard deviations near 0.15, about two and a half times that of Ridge. Adding weather to Ridge lowered R^2 to 0.916, raised MAE by 5.0% and raised RMSE by 1.5%. Random Forest, HistGradientBoosting and the weather ensembles also scored below historical Ridge. Weather added no usable signal (Table 4, Fig. 2). Area and previous production hold most of the information in 36 annual records, so the result applies to this setting. A weather effect on groundnut yield remains possible.

Table 4 Repeated cross-validation performance of the principal configurations, mean ± SD over 10×6 folds

Feature set	Model	R^2	MAE (t)	RMSE (t)
Historical	Ridge	0.920±0.058	3,923	5,298
Weather	Ridge	0.916±0.061	4,120	5,375
Historical	Gradient Boosting	0.838±0.150	6,319	10,442
Historical	Extra Trees	0.834±0.154	6,141	10,417

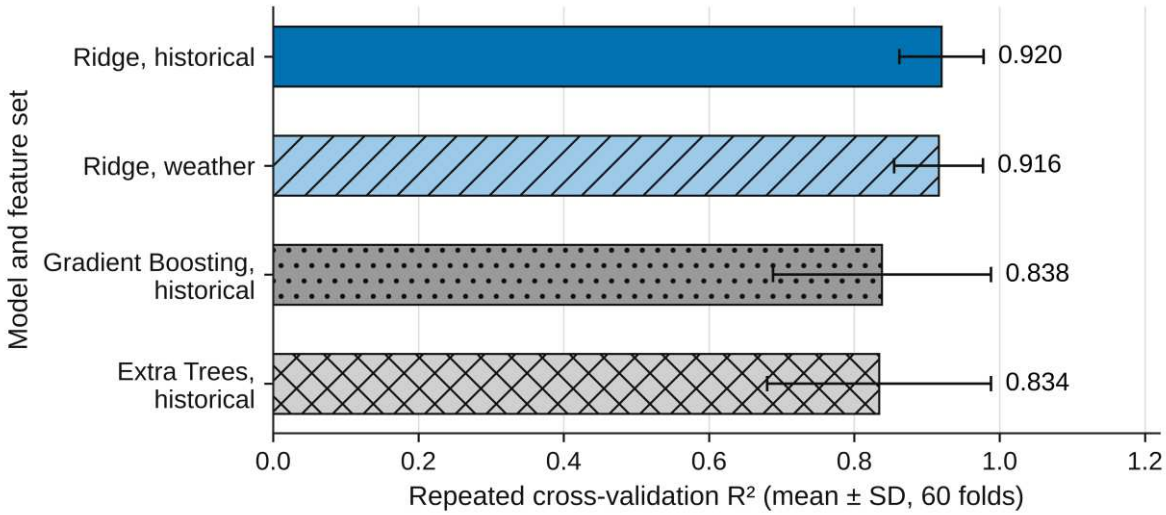


Fig. 2 Repeated cross-validation R^2 for the principal configurations, with bars for the mean and whiskers for one standard deviation over 60 folds

Previous production has a mean absolute SHAP value of 15,321 t, and cultivated area has 6,471 t. Previous production carries the persistence signal and area sets the scale, as Table 5 and Fig. 3 show. In a linear model the SHAP value of a feature is its coefficient times the feature's deviation from the mean, so these attributions are exact and restate the fitted coefficients. They describe how the surrogate uses its inputs and say nothing about crop-growth mechanisms.

Table 5 Global SHAP importance of the Ridge surrogate

Feature	Mean |SHAP| (t)	Role
Previous production	15,321	Persistence and state signal
Cultivated area	6,471	Production scale signal

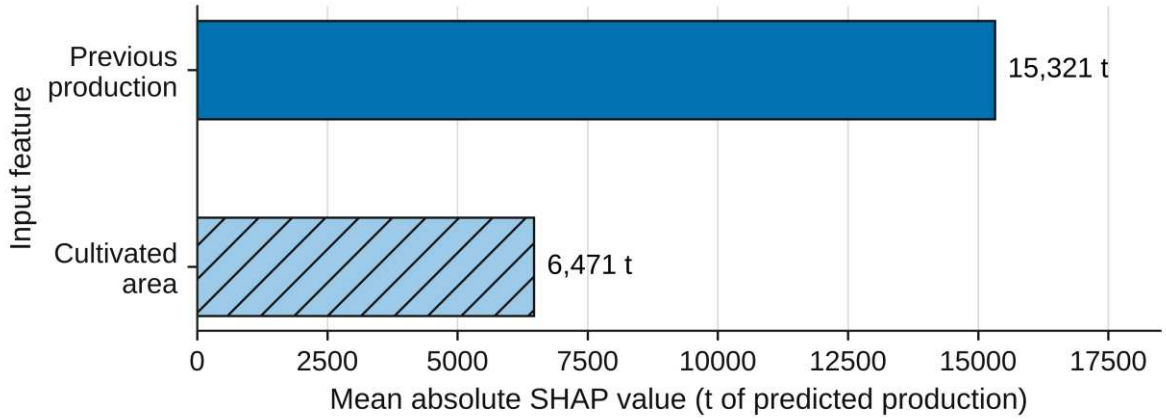


Fig. 3 Global SHAP importance for the Ridge production surrogate

5.2 Pareto optimization and balanced scenario

Each ablation moved the removed objective in the expected direction. Without the fertilizer objective, marketable food rose 6.1%, from 714,294 t to 757,581 t, and the fertilizer index rose from 0.842 to 0.936. Without the water objective, marketable food reached 724,697 t and water use 2.139×10^{9} m^3. Without the waste objective, waste rose from 3,989 t to 5,884 t. The full formulation scored a hypervolume of 0.39177 (Table 6, Fig. 4). Removing waste lowered the score by 16.5% to 0.32708, removing fertilizer by 3.2% to 0.37929, and removing water by 0.15% to 0.39119. Each formulation ran once, so the 0.15% water difference cannot be told apart from seed variation. Only the waste effect is large, and it holds for these response assumptions only.

Table 6 Objective values of one solution per formulation and normalized hypervolume

Formulation	Marketable food (t)	Water (m^3)	Fertilizer index	Waste (t)	Normalized hypervolume
Full 4-objective	714,294	2.046×10^{9}	0.842	3,989	0.39177
No waste	713,176	2.017×10^{9}	0.837	5,884	0.32708
No water	724,697	2.139×10^{9}	0.830	4,058	0.39119
No fertilizer	757,581	2.052×10^{9}	0.936	4,228	0.37929

Each row lists the balanced solution selected from the Pareto set of that formulation, so the decision vectors differ between rows.

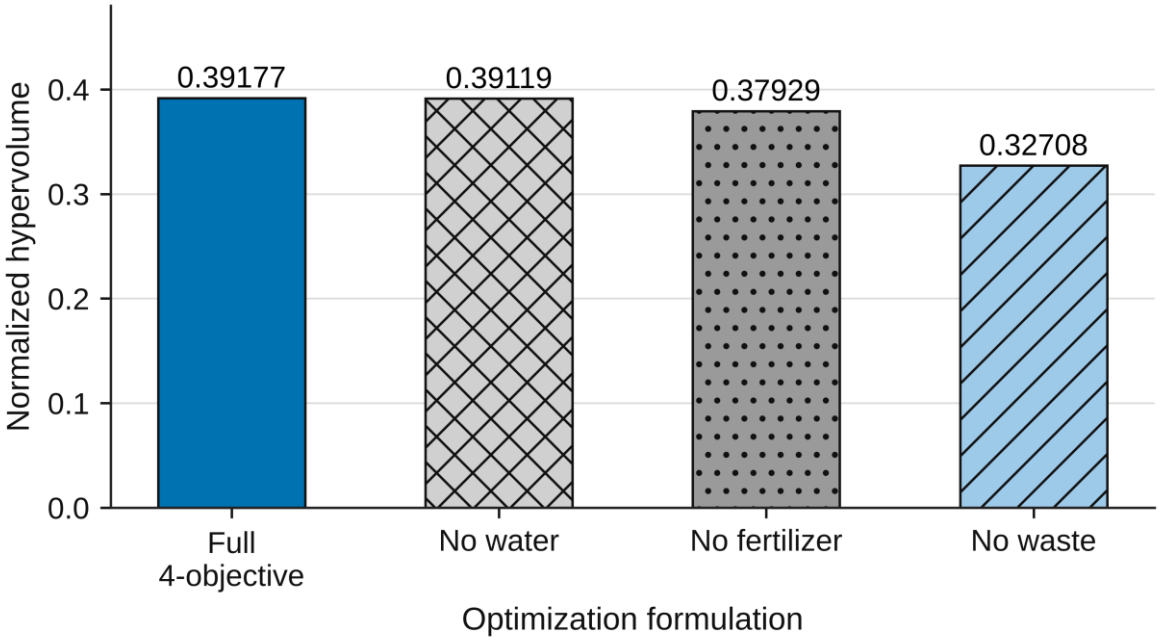


Fig. 4 Normalized hypervolume for the full formulation and the three objective ablations

The balanced solution of the full formulation lowers marketable food by 20.81%, water by 21.23%, the fertilizer-resource index by 15.81% and waste by 60.60% against the baseline (Table 7, Fig. 5). It gives up food to save resources and waste, and no solution in the set raises food while cutting every input.

Two derived quantities locate the source of these changes. Marketable food plus waste gives a harvested output of 718,283 t against 912,148 t at baseline, a fall of 21.25%, and water use fell by almost the same share, 21.23%. Food per cubic metre of water went from 0.347 to 0.349 kg, a gain of 0.5%. The optimizer saved water by producing less and found almost no gain in water productivity. The waste reduction comes from elsewhere: the selected solution implies a loss of 0.56% of harvested output, half of the 1.11% baseline and the largest cut Eq. 7 allows, which places the post-harvest control at its upper bound. Of the 6,136 t fall in waste, about 3,984 t, or 65%, follows from the lower loss rate and about 2,152 t, or 35%, from the smaller harvest.

Table 7 Baseline and balanced scenario

Metric	Baseline	Balanced scenario	Change
Marketable food	902,023 t[a]	714,294 t	−20.81%
Water use	2.598×10^9 m^3[a]	2.046×10^9 m^3	−21.23%
Fertilizer-resource index	1.000	0.842	−15.81%
Post-harvest waste	10,125 t	3,989 t	−60.60%
Harvested output[a]	912,148 t	718,283 t	−21.25%
Marketable food per m^3 of water[a]	0.347 kg	0.349 kg	+0.5%

[a] Derived from the reported values. Baseline marketable food is production minus the 1.11% loss, baseline water follows Eq. 8, and harvested output is marketable food plus waste.

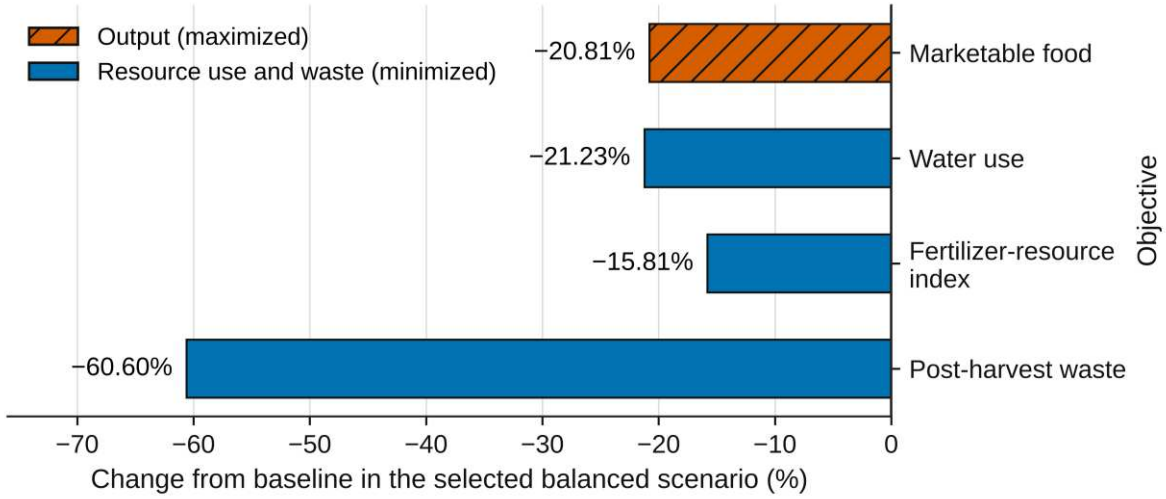


Fig. 5 Change from baseline in the balanced scenario

5.3 Sensitivity and uncertainty

Marketable production at the selected decision falls by 14,238 t, or 2.0%, for each 0.1 increase in K_y (Table 8, Fig. 6). From K_y = 0.50 to 0.90 the fall is 56,950 t, or 7.8% of the value at 0.50. The line is straight because we re-evaluated one fixed decision under Eq. 4 without re-running the optimizer. Raising the loss rate from 1.11% to 6.03% lifts waste from 3,989 t to 21,671 t, a factor of 5.43 that equals the ratio of the two rates. Waste scales one-to-one with the assumed loss rate.

Table 8 Sensitivity of the selected scenario to K_y and the post-harvest loss rate

Parameter	Value	Result	Change from main scenario
K_y	0.50	732,803 t marketable	+2.59%
K_y	0.60	718,565 t marketable	+0.60%
K_y	0.63, main	714,294 t marketable	0.00%
K_y	0.70	704,328 t marketable	−1.40%
K_y	0.80	690,091 t marketable	−3.39%
K_y	0.90	675,853 t marketable	−5.38%
Loss rate	1.11%, main	3,989 t waste	0.00%
Loss rate	6.03%	21,671 t waste	+443%

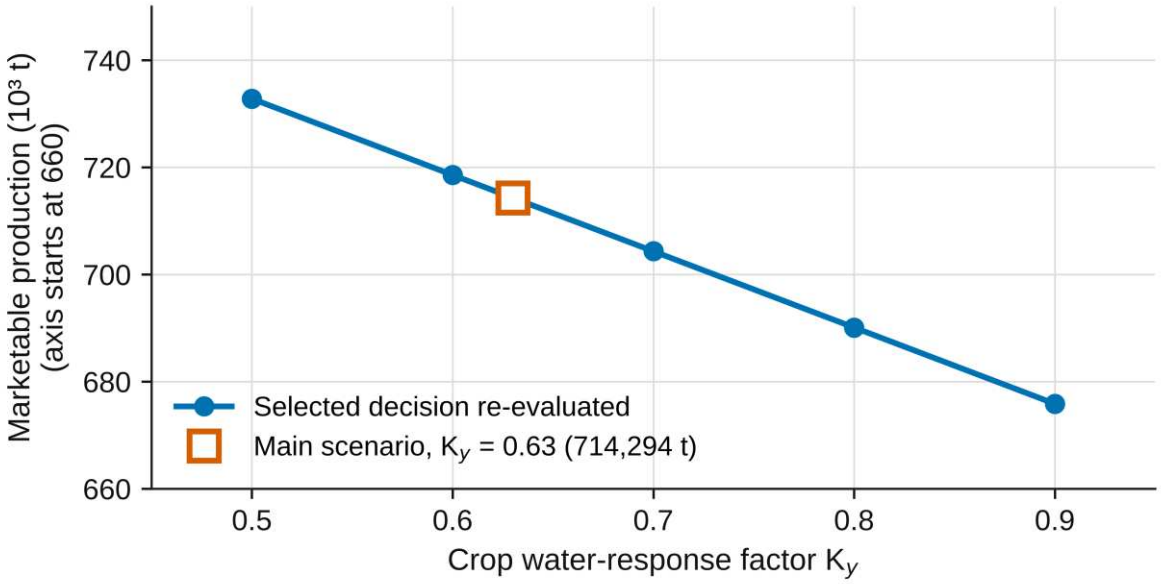


Fig. 6 Marketable production at the selected decision for different K_y values

A bootstrap over the 36 out-of-fold predictions gave a pooled R^2 of 0.957 with a 95% interval of 0.865–0.985, and Table 9 lists the error metrics. The pooled value exceeds the repeated-CV mean of 0.920 because the two statistics are computed differently. The pooled R^2 divides the error by the variance of all 36 districts. Each of the 60 fold-level R^2 values divides it by the variance of six districts, which is smaller and noisier. Repeated cross-validation remains the selection statistic.

Table 9 Bootstrap estimates on out-of-fold predictions

Metric	Estimate	95% interval
R^2	0.957	0.865–0.985
MAE	4,271 t	2,915–5,854 t
RMSE	6,214 t	4,092–8,441 t

6 Discussion

The surrogate leans on year-to-year persistence. Previous production dominates its attributions, and weather variables did not improve it. A multi-year panel should come before richer climate features.

The twin adds what the surrogate cannot give. A production model does not say whether a higher-output scenario is worth the extra water or the extra post-harvest loss. Passing the predicted state through explicit resource and loss responses turns one estimate into a set of alternatives that a planner can rank by policy priorities. Purcell and Neubauer name this use of what-if simulation as underdeveloped in agricultural twins [5].

The waste ablation produced the largest change in hypervolume, and the decomposition in Section 5.2 qualifies the headline numbers. The 21.23% water saving is not an efficiency gain, because output fell by nearly the same share. The 60.60% waste cut follows from a control variable at its bound and a 50% maximum

effect that we assumed. The balanced solution illustrates a trade-off. It is not advice for a farm.

Ridge beat the tree ensembles on 36 observations, and no larger panel was available to test that order.

7 Limitations and responsible use

The data are annual and district-level, so they miss field variation, within-season dynamics and sensor readings. The totals check covers 2022–23 area and production only. Cross-validation split districts at random, and neighbouring districts share climate and management. Correlated errors inflate scores in that case. No spatially blocked or leave-region-out split [29] was run, so generalization to unseen districts is untested. No persistence baseline that copies previous production was evaluated, so the gain of the fitted model over that rule is unknown. The weather summaries cover one season per district and miss crop stages. Geocoded district points can fall outside the groundnut-growing area.

The fertilizer response is a bounded scenario function. The water and loss parameters come from the literature and need field calibration. Each formulation ran once with seed 42, so the hypervolume differences are single-run values. No regressor had its hyperparameters tuned, and tuning could change the ranking. Local SHAP explanations for single districts remain for future work.

AgriFoodTwin-XAI is a decision-support prototype. It prescribes nothing to farmers. Any deployment needs local agronomic review, field validation and a clear statement of uncertainty. Held-out observations validate the production model. Resource use and waste changes are counterfactual outputs of the twin.

8 Conclusion

This study built a computational digital twin for groundnut in 36 Tamil Nadu districts that links a production surrogate, a SHAP explanation, a scenario model and a four-objective NSGA-II search. A Ridge surrogate using cultivated area and previous production predicted district production with R^2 of 0.920±0.058 over 60 repeated folds, and adding weather variables did not improve it. SHAP ranked previous production above area, with mean absolute values of 15,321 t and 6,471 t. In the optimization, the full four-objective formulation reached a normalized hypervolume of 0.39177, against 0.32708 without the waste objective. The balanced solution cut modelled water use by 21.23% and post-harvest waste by 60.60% at a cost of 20.81% of marketable food. The decomposition shows where these changes come from. The water saving follows the 21.25% fall in harvested output and gives almost no gain in water productivity. The waste saving comes from the assumed halving of the loss rate for 65% of the cut and from the smaller harvest for the rest. All three reductions are simulated and rest on assumed response functions, so they are not field results. The evidence also comes from one state, one crop, one year and 36 records, and validation used random folds rather than spatial blocks, with no

persistence baseline and a single random seed for the optimizer. The next step is to calibrate the response parameters with measured irrigation, fertilizer and post-harvest data collected over several seasons, and to repeat the validation with spatial cross-validation and a persistence baseline before the twin informs any planning decision. Local SHAP explanations for individual districts would also show where the surrogate departs from the statewide pattern.

Declarations

Funding. The authors received no funding for the research, authorship, or publication of this article.

Competing interests. The authors declare that they have no competing interests.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Clinical trial number. Not applicable.

Data availability. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Author contributions. Arshed Ayoob P: Data curation, Software, Investigation, Visualization, Writing – original draft. Yoganand S: Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing – review and editing. Arun Pandian J: Conceptualization, Methodology, Formal analysis, Validation, Writing – review and editing, Supervision. All authors read and approved the final manuscript.

Acknowledgements. The authors gratefully acknowledge Vellore Institute of Technology (VIT), Vellore, India, for providing institutional support towards the article processing charge (APC), as well as the infrastructure and research environment necessary to conduct this study.

References

[1] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: a systematic literature review,” Comput. Electron. Agric., vol. 177, 105709, 2020, https://doi.org/10.1016/j.compag.2020.105709

[2] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, “Machine learning in agriculture: a comprehensive updated review,” Sensors, vol. 21, no. 11, 3758, 2021, https://doi.org/10.3390/s21113758

[3] C. Verdouw, B. Tekinerdogan, A. Beulens, and S. Wolfert, “Digital twins in smart farming,” Agric. Syst., vol. 189, 103046, 2021, https://doi.org/10.1016/j.agsy.2020.103046

[4] C. Pylianidis, S. Osinga, and I. N. Athanasiadis, “Introducing digital twins to agriculture,” Comput. Electron. Agric., vol. 184, 105942, 2021, https://doi.org/10.1016/j.compag.2020.105942

[5] W. Purcell and T. Neubauer, “Digital twins in agriculture: a state-of-the-art review,” Smart Agric. Technol., vol. 3, 100094, 2023, https://doi.org/10.1016/j.atech.2022.100094

[6] N. Peladarinos, D. Piromalis, V. Cheimaras, E. Tserepas, R. A. Munteanu, and P. Papageorgas, “Enhancing smart agriculture by implementing digital twins: a comprehensive review,” Sensors, vol. 23, no. 16, 7128, 2023, https://doi.org/10.3390/s23167128

[7] W. Purcell, T. Neubauer, and K. Mallinger, “Digital twins in agriculture: challenges and opportunities for environmental sustainability,” Curr. Opin. Environ. Sustain., vol. 61, 101252, 2023, https://doi.org/10.1016/j.cosust.2022.101252

[8] K. M. Krithika, N. Maheswari, and M. Sivagami, “Models for feature selection and efficient crop yield prediction in the groundnut production,” Res. Agric. Eng., vol. 68, no. 3, pp. 131–141, 2022, https://doi.org/10.17221/15/2021-RAE

[9] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam, “A systematic literature review on crop yield prediction with deep learning and remote sensing,” Remote Sens., vol. 14, no. 9, 1990, 2022, https://doi.org/10.3390/rs14091990

[10] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?,” in Adv. Neural Inf. Process. Syst. 35 (NeurIPS 2022), Datasets and Benchmarks Track, 2022.

[11] A. E. Hoerl and R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970, https://doi.org/10.1080/00401706.1970.10488634

[12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 4765–4774.

[13] S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nat. Mach. Intell., vol. 2, pp. 56–67, 2020, https://doi.org/10.1038/s42256-019-0138-9

[14] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat. Mach. Intell., vol. 1, pp. 206–215, 2019, https://doi.org/10.1038/s42256-019-0048-x

[15] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, 2002, https://doi.org/10.1109/4235.996017

[16] F. Chen, N. Cui, S. Jiang, et al., “Multi-objective deficit drip irrigation optimization of citrus yield, fruit quality and water use efficiency using NSGA-II in seasonal arid area of Southwest China,” Agric. Water Manag., vol. 287, 108440, 2023, https://doi.org/10.1016/j.agwat.2023.108440

[17] X. Liu and D. Yang, “Irrigation schedule analysis and optimization under the different combination of P and ET0 using a spatially distributed crop model,” Agric. Water Manag., vol. 256, 107084, 2021, https://doi.org/10.1016/j.agwat.2021.107084

[18] Y. Wang, W. Shi, and T. Wen, “Prediction of winter wheat yield and dry matter in North China Plain using machine learning algorithms for optimal water and nitrogen application,” Agric. Water Manag., vol. 277, 108140, 2023, https://doi.org/10.1016/j.agwat.2023.108140

[19] J. Doorenbos and A. H. Kassam, Yield Response to Water, FAO Irrigation and Drainage Paper 33. Rome, Italy: FAO, 1979.

[20] P. Steduto, T. C. Hsiao, E. Fereres, and D. Raes, Crop Yield Response to Water, FAO Irrigation and Drainage Paper 66. Rome, Italy: FAO, 2012.

[21] P. Steduto, T. C. Hsiao, D. Raes, and E. Fereres, “AquaCrop: the FAO crop model to simulate yield response to water. I. Concepts and underlying principles,” Agron. J., vol. 101, no. 3, pp. 426–437, 2009, https://doi.org/10.2134/agronj2008.0139s

[22] NABCONS, Study to Determine Post-Harvest Losses of Agri Produces in India. Mumbai, India: NABARD Consultancy Services, 2022.

[23] FAO, The State of Food and Agriculture 2019: Moving Forward on Food Loss and Waste Reduction. Rome, Italy: FAO, 2019.

[24] J. Parfitt, M. Barthel, and S. Macnaughton, “Food waste within food supply chains: quantification and potential for change to 2050,” Philos. Trans. R. Soc. B, vol. 365, no. 1554, pp. 3065–3081, 2010, https://doi.org/10.1098/rstb.2010.0126

[25] Department of Economics and Statistics, Government of Tamil Nadu, Statistical Handbook of Tamil Nadu 2022–23. Chennai, India: Department of Economics and Statistics, Government of Tamil Nadu.

[26] NASA POWER Project, “Prediction Of Worldwide Energy Resources (POWER) data access viewer,” NASA Langley Research Center. [Online]. Available: https://power.larc.nasa.gov.

[27] F. Pedregosa et al., “Scikit-learn: machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[28] E. Zitzler and L. Thiele, “Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach,” IEEE Trans. Evol. Comput., vol. 3, no. 4, pp. 257–271, 1999, https://doi.org/10.1109/4235.797969

[29] D. R. Roberts et al., “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” Ecography, vol. 40, no. 8, pp. 913–929, 2017, https://doi.org/10.1111/ecog.02881