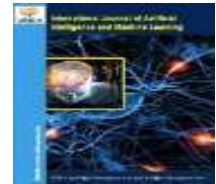




## International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

# AgroConform: Coverage-Guaranteed Crop Recommendation with Sensor-Noise-Aware Calibration, Distribution-Shift Monitoring and Cost-Aware Agronomic Recourse

Prasun Chowdhury<sup>1\*</sup>

<sup>1</sup>Department of Electronics and Communication Engineering, St. Thomas' College of Engineering and Technology, Kolkata, West Bengal, India. E-mail: [prasun.chowdhury@stcet.ac.in](mailto:prasun.chowdhury@stcet.ac.in)

### Abstract

Crop recommendation models now exceed 99% accuracy on standard soil and climate data, yet a single predicted crop does not show how far that advice can be trusted in a given field. This study presents AgroConform, which wraps any probabilistic classifier with split conformal prediction and maps each output to one of three actions: recommend, shortlist or abstain. On 2,200 records covering 22 crops, the framework met its 95% coverage target, and recommended crops were correct in 99.6% of cases. Noise-aware calibration held coverage at 95% under simulated soil-sensor error, whereas standard calibration fell to 66%. A shift monitor flagged 49% of fields after a simulated 6 °C warming, while the classifier changed its advice for fewer than 1%. A cost-aware recourse search priced the soil amendments that make a preferred crop admissible and showed that most refusals are climatic. Each field takes about 15 ms on a standard CPU.

**Keywords:** Conformal prediction; Crop recommendation; Uncertainty quantification; Distribution shift; Counterfactual recourse; Precision agriculture; Trustworthy machine learning

## 1. Introduction

### 1.1 Background and motivation

Choosing what to sow is one of the few decisions in a cropping cycle that cannot be revisited cheaply. Once seed, fertiliser and labour are committed, a poor match between crop and field is paid for over a whole season. Data-driven crop recommendation has therefore drawn steady interest, helped by cheaper soil probes, village-level weather stations and the spread of Internet of Things (IoT) platforms in farming [1, 2]. Reviews describe a crowded field of classifiers and regressors trained on soil nutrients, weather variables and remote-sensing indices [3, 4], with tree ensembles and neural networks used most often for both crop choice and yield [5, 6].

The topic is timely for three reasons. First, crop recommenders are moving out of notebooks and into the field: recent systems read N, P, K, pH and weather from IoT nodes and run the model on an edge device next to the farm [7]. Second, the probes feeding these systems are inexpensive and imperfect. A field study of a portable soil nutrient detector in Indonesian paddy soils found that its N, P and K readings shifted strongly with soil moisture, and noted that no accuracy assessment against laboratory values was yet available [8]. Third, the climate in which a model is used increasingly differs from the climate in which it was trained, and warming already measurably affects crop performance [9, 10]. A recommender that is accurate on its own test split may therefore be used on noisy inputs, under unfamiliar weather, by people who cannot see when it is guessing.

On the widely used benchmark of 2,200 soil and climate records covering 22 crops, accuracy no longer separates good methods from weak ones. Random forests [11], gradient-boosted trees [12, 13] and support vector machines all exceed 97%, and the best exceed 99%. Many studies stop there and add a feature-attribution chart [14, 15]. Three practical questions remain open. How far can a given recommendation be trusted for a given field? What happens when inputs come from cheap sensors or from an unusual season? And what can a farmer do when the recommended crop is not the one he or she wants to grow?

Explanation methods alone do not answer these questions. An attribution profile describes the model but does not bound its error rate or say when to stop trusting it [16, 17]. Class probabilities are a natural starting point, but tree ensembles and kernel machines are often miscalibrated [18, 19], and Bayesian or ensemble estimates of uncertainty carry no finite-sample guarantee [20]. Conformal prediction closes this gap. Given an exchangeable calibration sample, it turns any model score into prediction sets that contain

the true label with a probability chosen by the user, without assumptions about the model or the data [21, 22].

### 1.2 Related work

Machine learning in agriculture. Liakos et al. [4] and Kamilaris and Prenafeta-Boldú [3] catalogue supervised and deep learning across crop, livestock, water and soil applications, and note that most crop-level studies are evaluated on random hold-out splits from a single source. Jeong et al. [5] found random forests competitive with process-based models for global and regional yield, and Khaki and Wang [6] used deep networks with genotype and weather inputs. Crop recommendation studies built on soil nutrients and weather usually report accuracy, precision, recall and F1, sometimes with SHAP or LIME charts, but rarely calibration, coverage or behaviour under changed conditions.

Calibration and uncertainty. Guo et al. [18] showed that modern classifiers can be confidently wrong and proposed temperature scaling, while Naeini et al. [19] introduced Bayesian binning and the expected calibration error (ECE) used in this study. Abdar et al. [20] survey Bayesian, ensemble and evidential approaches. These methods improve probabilities but do not provide a statement such as “the correct crop lies in this set at least 95% of the time”.

Conformal prediction. The split or inductive form [23, 24] needs one model fit and a held-out calibration sample. For classification, Sadinle et al. [25] derived the least-ambiguous set-valued classifier (LAC), Romano et al. [26] proposed adaptive prediction sets (APS), and Angelopoulos et al. [27] added regularisation (RAPS) to keep sets small. Vovk [28] analysed class-conditional or Mondrian coverage. Beyond exchangeability, Tibshirani et al. [29] reweighted calibration scores under covariate shift, Barber et al. [30] bounded the coverage gap under general non-exchangeability, and Gibbs and Candès [31] proposed online adaptation; covariate shift itself has a long history [32–34]. In agriculture, conformal methods have so far been applied mainly to image-based crop and weed classification, where Melki et al. [35] showed that marginal guarantees can fail for groups defined by pedo-climatic conditions, and Melki et al. [36] used them to bound errors in automated precision weeding. Hullman et al. [37] examine how prediction sets should be presented to people who must act on them. We have not found a study that applies conformal prediction to tabular soil and climate crop recommendation, or combines it with sensor noise, climatic drift and actionable recourse.

Counterfactual explanations and recourse. Wachter et al. [38] defined counterfactual explanations as the smallest input change that alters a decision. Ustun et al. [39] restricted such changes to variables a person can control and attached a cost to each, Mothilal et al. [40] generated diverse counterfactuals, and Karimi et al. [41] separate explaining a model from recommending a feasible action. Agriculture suits this view well: nutrients and pH can be amended at a known cost, whereas temperature and rainfall cannot.

### 1.3 Contributions

This article makes five contributions.

- It recasts crop recommendation as set-valued prediction with a three-way rule (recommend, shortlist, abstain) and compares five set constructions over 30 random splits and four base learners, reporting both average and worst-crop coverage with statistical tests.
- It shows that the coverage guarantee breaks quickly under realistic soil-sensor error and proposes noise-aware calibration, which restores nominal coverage when the noise level is roughly known, and it measures the cost of under- or over-estimating that level.
- It introduces a shift monitor that pairs a nearest-neighbour conformal p-value with weighted conformal prediction [29] and reports the effective sample size of the weights, so users can see when a weighted guarantee has become uninformative.
- It formulates cost-aware agronomic recourse, in which only soil nutrients and pH may change and climate is fixed, and shows that most failed recourse queries are explained by climate rather than soil.
- It proposes coverage-drop permutation importance, which ranks inputs by how much the reliability of the prediction sets depends on them, and translates all findings into deployment guidance for extension services.

Figure 1 outlines the framework. Section 2 describes the data and methods, Section 3 reports the results, Section 4 discusses practical implications and limitations, and Section 5 concludes.

## 2. Materials and Methods

### 2.1 Data

The experiments use the public crop recommendation dataset (see Data Availability). It holds 2,200 records with seven attributes: soil nitrogen (N), phosphorus (P) and potassium (K), soil pH, air temperature, relative humidity and rainfall. Each of the 22 crops has exactly 100 records. Table 1 summarises the attributes and Figure 2 shows the mean of each attribute for each crop. Several crops overlap strongly: rice and jute share warm, humid, high-rainfall conditions, and the pulses mothbean, mungbean and blackgram

differ mainly in humidity and rainfall. No records were removed. Because the classes are balanced, coverage can be estimated with equal precision for every crop.

## 2.2 Base learners and evaluation protocol

Six classifiers were trained with scikit-learn [42]: multinomial logistic regression (LR), distance-weighted k-nearest neighbours (kNN,  $k = 7$ ), a support vector machine with an RBF kernel and Platt-scaled probabilities (SVM-RBF), a two-layer perceptron (MLP, 64 and 64 units), histogram gradient boosting (HistGB, 300 iterations, learning rate 0.05) and a random forest (RF, 300 trees). Distance-based and neural models received standardised inputs. Point performance was estimated by stratified five-fold cross-validation using accuracy, macro-F1, negative log-likelihood (NLL), the multi-class Brier score and the 15-bin ECE [19]. Unless stated otherwise, every conformal experiment used a stratified 50/25/25 split into training, calibration and test sets, repeated over 30 random seeds. Differences in worst-crop coverage between methods were tested with the paired Wilcoxon signed-rank test over these 30 splits, and 95% confidence intervals for mean coverage were computed with the t distribution. Table 2 lists the main symbols.

## 2.3 Split conformal prediction for crops

Let  $\hat{f}$  be a classifier trained on the training set that returns class probabilities  $\hat{p}(k | x)$  for crops  $k = 1, \dots, K$ . For each of the  $n$  calibration pairs  $(x_i, y_i)$ , a non-conformity score  $s_i = s(x_i, y_i)$  is computed, and the threshold

$$\hat{q} = \text{the } [(n + 1)(1 - \alpha)]\text{-th smallest value of } \{s_1, \dots, s_n\} \cup \{1\}$$

is stored. For a new field  $x$ , the prediction set is  $C(x) = \{k : s(x, k) \leq \hat{q}\}$ . If calibration and test pairs are exchangeable,  $P(y \in C(x)) \geq 1 - \alpha$  for any classifier and any data distribution [21, 23]. Three scores were compared. LAC uses  $s(x, k) = 1 - \hat{p}(k | x)$  [25]. APS sorts the probabilities in decreasing order and uses the randomised cumulative probability up to and including class  $k$  [26]. RAPS adds a penalty  $\lambda \cdot \max(0, r(k) - k_{\text{reg}})$  to the APS score, where  $r(k)$  is the rank of class  $k$ , with  $\lambda = 0.02$  and  $k_{\text{reg}} = 2$  [27]. Mondrian LAC computes a separate threshold  $\hat{q}_k$  from the calibration records of each crop; this guarantees coverage within every crop, provided each crop has at least  $\lceil 1/\alpha \rceil - 1$  calibration records [28]. As a baseline without calibration, the naive cumulative-probability set adds classes in decreasing order of probability until their total reaches  $1 - \alpha$ .

In plain terms, the calibration step asks how confident the model had to be, on fields it had not seen, to include the correct crop 95% of the time. The answer,  $\hat{q}$ , then sets the bar for every new field: any crop that clears it goes into the set.

## 2.4 Three-way advisory rule

The size of the set carries information that a single label hides. AgroConform maps each field to one of three actions. If  $C(x)$  holds exactly one crop and the shift monitor (Section 2.6) does not flag  $x$ , that crop is recommended. If  $C(x)$  holds two or more crops, they are offered as a shortlist, and the farmer chooses using local knowledge, prices or rotation needs. If  $C(x)$  is empty, or the monitor flags  $x$  as atypical, the system abstains and refers the case to an extension officer. An empty LAC set means that no crop reaches the confidence that, on calibration data, separates correct labels from the rest. Treating it as abstention rather than as an error follows the bounded-error view of Sadinle et al. [25].

## 2.5 Noise-aware calibration for low-cost soil sensors

In the field, N, P, K and pH are increasingly read by inexpensive electrochemical or optical probes whose errors are far larger than those of laboratory analysis and which respond to soil moisture [8]. The training data, by contrast, reflect cleaner measurements. If test inputs are noisy but calibration inputs are not, exchangeability fails and the guarantee no longer holds. A field reading was therefore modelled as  $\tilde{x}_j = \max(0, x_j + \sigma \cdot \text{SD}_j \cdot \varepsilon_j)$  for the four soil attributes  $j$ , with  $\varepsilon_j$  drawn from a standard normal distribution and  $\text{SD}_j$  the standard deviation of attribute  $j$ ; weather attributes were left unchanged. Noise-aware calibration replaces each calibration record with three independently perturbed copies at an assumed noise level  $\hat{\sigma}$  before computing  $\hat{q}$ . When  $\hat{\sigma} = \sigma$ , the perturbed calibration scores are exchangeable with the test scores and the guarantee is restored. Noise levels  $\sigma$  from 0 to 0.5 and assumed levels  $\hat{\sigma} \in \{0, 0.5\sigma, \sigma, 2\sigma\}$  were evaluated to measure the effect of getting the noise level wrong.

## 2.6 Shift monitor

The monitor has two parts. The first detects atypical inputs. A nearest-neighbour score  $d(x)$ , the mean Euclidean distance from standardised  $x$  to its five nearest training records, is computed for calibration and test fields, and the conformal p-value

$$p(x) = [1 + \#\{i : d(x_i) \geq d(x)\}] / (n + 1) \quad (2)$$

is uniformly distributed when the field resembles the calibration data. A field is flagged when  $p(x) \leq 0.05$  (or 0.01 inside the three-way rule), which holds the false-alarm rate at that level when no shift is present.

The second part handles a known covariate shift. Following Tibshirani et al. [29], each calibration score receives a weight  $w(x_i)$  equal to the estimated density ratio between target and source inputs, obtained from a logistic-regression classifier that separates calibration records from an unlabelled sample of target fields. The threshold for test point  $x$  becomes the  $(1 - \alpha)$  quantile of the weighted distribution that places mass  $w(x_i)/[\sum w(x_i) + w(x)]$  on  $s_i$  and the remaining mass on  $+\infty$ . Weighted guarantees weaken when a few weights dominate, so the monitor also reports the effective sample size

$$ESS = (\sum_i w_i)^2 / \sum_i w_i^2 \quad (3)$$

of the calibration weights. To create labelled shifted test sets, test fields were resampled with probability proportional to  $\exp(\lambda \cdot g(x))$ , where  $g$  is a standardised index for one of three agronomic scenarios: hot and dry (temperature minus rainfall), cool and humid (humidity minus temperature), and acidic, low-potassium soil (negative pH minus K). Larger  $\lambda$  means a stronger shift, and  $\lambda = 0$  reproduces the unshifted case. Each scenario and  $\lambda$  value was repeated 15 times with 880 training, 440 calibration and 880 pool records.

To study climatic drift without labels, temperature was raised by 0–6 °C and rainfall reduced by 0–60% on the test fields of a fixed split. Because the correct crop is then unknown, the analysis reports the share of fields with empty sets, the share flagged by the monitor and the share whose top-ranked crop changed.

### 2.7 Cost-aware agronomic recourse

Suppose a farmer wishes to grow crop  $t$  but  $t \notin C(x)$ . The recourse engine looks for a modified field  $x'$  that minimises the cost

$$c(x, x') = \sum_j \pi_j \cdot \max(0, x'_j - x_j) + \pi^+ \cdot \max(0, \Delta) + \pi^- \cdot \max(0, -\Delta) \quad (4)$$

where  $j$  runs over N, P and K,  $\Delta = x'_pH - x_pH$  is the change in pH, and  $\pi^+$  and  $\pi^-$  are the unit costs of raising and lowering pH. The search is constrained so that  $t \in C(x')$ , and temperature, humidity and rainfall stay at their observed values. Nutrients may be added but not removed, because depleting soil nutrients is not possible within one season. The unit costs are indicative and can be changed by the user: INR 13 per kg N, INR 59 per kg  $P_2O_5$  and INR 55 per kg  $K_2O$ , broadly in line with subsidised retail fertiliser prices in India, and INR 2,500 (liming) or INR 4,000 (acidifying amendment) per 0.1-unit pH change per hectare. The random forest is not differentiable, so the search is guided by prototypes. For every training record of crop  $t$ , candidate fields are generated by moving  $x$  a fraction  $\tau \in \{0.25, 0.5, 0.75, 1\}$  of the way toward that record in the adjustable attributes (upward only for N, P and K). All candidates are scored in one batch, and the cheapest one that places  $t$  in the conformal set is returned. A query is called climate-compatible when the field's temperature, humidity and rainfall fall inside the range observed for crop  $t$  in the training data, widened by 5% of a standard deviation. Six test fields per crop were paired with each of the 21 alternative crops, giving 2,772 queries.

### 2.8 Coverage-drop permutation importance

Classical permutation importance measures the fall in accuracy when one attribute is shuffled [43]. For a system whose output is a set, the more useful question is how much the reliability of that set depends on each attribute. The coverage drop of attribute  $j$  is therefore defined as the coverage of  $C$  on the original test set minus its coverage on a copy in which attribute  $j$  is randomly permuted, with  $\hat{q}$  held fixed. Computing the same quantity for each crop gives a crop-by-attribute map of where the guarantee relies on which measurement. Each value is averaged over 20 permutations.

### 2.9 Implementation

Table 3 summarises the complete procedure. All code was written in Python 3 with scikit-learn and NumPy and run on a two-core CPU. Training the random forest took 0.6 s. Serving one field, including the conformal set and the monitor's p-value, took about 15 ms on a single core, and about 0.05 ms per field when fields were processed in batches. Recourse queries averaged under 2 ms each in batched mode. The framework is therefore light enough for an edge gateway or a low-cost advisory server.

## 3. Results

### 3.1 Point accuracy and calibration

Table 4 and Figure 3 compare the six learners. The random forest was the most accurate at  $99.50 \pm 0.30\%$ , followed by HistGB (98.95%) and SVM-RBF (98.86%); every model exceeded 97%. Accuracy and calibration did not go together. HistGB and the MLP had the lowest ECE (0.70% and 0.99%) and the lowest NLL. The random forest, despite its accuracy, had an ECE of 4.13%, and SVM-RBF was the worst calibrated at 11.06%. The reliability curves in Figure 3(c) show that both RF and SVM-RBF sit above the diagonal: they are under-confident and assign probabilities of about 0.6–0.8 to predictions that are almost always right.

This matters for LAC, because an under-confident model produces conformal thresholds that exclude many correct predictions, as the next subsection shows.

### 3.2 Average and crop-wise coverage

Table 5 and Figure 4 report the conformal results with the random forest. Every conformal method met its target coverage within sampling error. At  $\alpha = 0.05$ , mean coverage was 0.955 for LAC (95% CI 0.950–0.960), 0.950 for APS and RAPS (0.946–0.953) and 0.964 for Mondrian LAC (0.959–0.968). The naive cumulative-probability set covered every test field, but only by returning 1.59 crops on average against 0.96 for LAC, which is the price of skipping calibration on an under-confident model.

Because the forest is so accurate, LAC sets at  $\alpha \geq 0.02$  never hold more than one crop. Coverage is reached by leaving out the least confident fields, which receive empty sets (4.4% at  $\alpha = 0.05$ ; Figure 4c). At  $\alpha = 0.01$  the threshold rises, empty sets fall to 0.3%, and LAC remains the most compact construction with 1.00 crops per set against 1.82 for APS and 1.55 for RAPS. Mondrian LAC cannot be used at  $\alpha \leq 0.02$  with 25 calibration records per crop: the required rank exceeds the sample size, so every set becomes the full list of 22 crops. This gives a concrete sample-size rule for anyone who needs per-crop guarantees.

Average coverage hides large differences between crops (Figure 5). With LAC at  $\alpha = 0.05$ , jute was covered in only 79.5% of cases, mothbeans and rice in about 82%, and blackgram and lentil in about 91%, while apple, grapes and muskmelon were covered in every split. These are the crops whose climatic ranges overlap with a neighbour's. Mean worst-crop coverage rose from 0.732 with LAC to 0.848 with APS and RAPS (paired Wilcoxon test over 30 splits,  $p < 0.001$ ) and to 0.816 with Mondrian LAC ( $p < 0.001$ ). APS spends set size where the model is uncertain and so protects the difficult crops, whereas LAC spreads it evenly. Table 6 shows the same pattern with HistGB, SVM-RBF and kNN as base learners, so the finding is not specific to the forest. In practice, a service that cares about every crop, not just the average field, should prefer APS or RAPS.

### 3.3 Three-way advisory decisions

Table 7 converts the LAC sets into advisory actions, with the monitor threshold at  $p \leq 0.01$ . At  $\alpha = 0.01$ , AgroConform recommended a single crop for 98.6% of test fields, offered a shortlist for 0.3% and abstained on 1.1%. The recommended crop was correct in 99.62% of cases. Raising  $\alpha$  to 0.05 increased abstention to 5.2% and lifted the accuracy of recommended fields to 99.88%; at  $\alpha = 0.10$  accuracy reached 99.98% with 9.9% abstention. In one representative split at  $\alpha = 0.05$ , the 32 abstained fields had a top-ranked accuracy of 84.4%, against 99.9% for recommended fields. Abstention therefore concentrates errors: about one abstained field in six would have received the wrong crop, against roughly one in a thousand among recommended fields. The level  $\alpha$  gives an extension service a single, interpretable control over how much is automated and how much is referred to a person.

### 3.4 Covariate shift and the limits of reweighting

Figure 6 shows how standard and weighted conformal prediction behave as the test population drifts away from the calibration population. The outcome depends on the direction of the shift. Under hot, dry shift, standard sets stayed above 0.95, because the shifted population is dominated by well-separated crops. Under cool, humid shift, coverage rose to 0.997, since humid fields contain crops the forest assigns with high confidence. Under the acidic, low-potassium scenario, standard coverage fell steadily to 0.900 at  $\lambda = 2.5$ . Top-ranked accuracy stayed above 99% in this scenario, so the loss did not come from misclassification but from empty sets: the forest is less confident in this part of the input space, and a threshold calibrated on the original population excludes those fields.

Weighted conformal prediction restored coverage in the acidic scenario (0.975 at  $\lambda = 2.5$ ) and kept it close to 0.95 in the cool, humid case at moderate  $\lambda$ . The cost appears in the lower row of Figure 6. The effective sample size of the 440 calibration weights fell from 419 at  $\lambda = 0$  to below 120 at  $\lambda = 1$ , and to 5–8 at  $\lambda = 2.5$ . In that regime the weighted quantile is often infinite, and mean set size grows to 17–18 of the 22 crops, a guarantee that is valid but useless. In these runs an ESS below about 100 coincided with set sizes above 1.9. That is a sensible point to stop relying on weighted sets and switch to abstention or to recalibration with local data.

### 3.5 Soil-sensor noise and noise-aware calibration

Sensor noise was the most damaging departure studied (Figure 7). With clean calibration, coverage fell from 0.953 without noise to 0.898 at  $\sigma = 0.05$ , 0.661 at  $\sigma = 0.2$  and 0.384 at  $\sigma = 0.5$ , even though top-ranked accuracy at  $\sigma = 0.2$  was still 95.3%. Sets shrank as noise grew (mean size 0.39 at  $\sigma = 0.5$ ), because noisy inputs push the forest's confidence below the clean threshold and most fields receive empty sets. A system that presented these sets as guaranteed would be seriously misleading.

Noise-aware calibration at the correct level held coverage between 0.953 and 0.958 at every  $\sigma$ . Sets grew only as much as needed: 1.01 crops at  $\sigma = 0.2$  and 1.96 crops at  $\sigma = 0.5$ , which is the honest response to readings that can no longer separate similar crops. Errors in the assumed noise level had unequal effects. Assuming half the true noise left coverage at 0.84–0.93, whereas assuming twice the true noise over-

covered at about 0.99 with sets of up to 3.41 crops. In practice, it is safer to take a slightly pessimistic noise level from the probe's data sheet or from a few paired laboratory tests than an optimistic one.

### 3.6 Silent failure under climatic drift

Figure 8 reports the climatic perturbations. When temperature was raised by up to 6 °C, the forest changed its top-ranked crop for fewer than 1% of fields. The model is effectively blind to temperature beyond the training range and would keep recommending the same crop in a much warmer season with no sign of doubt. The monitor, by contrast, flagged 20.2% of fields at +4 °C and 48.9% at +6 °C, and empty LAC sets rose from 5.3% to 26.0%. Rainfall deficits produced a different pattern: at a 40% deficit the top-ranked crop changed for 12.5% of fields, empty sets rose to 54.9% and the monitor flagged 13.1%, rising to 37.6% at a 60% deficit. Each signal catches something the other misses. The distance-based flag responds most to warming, which the classifier ignores, while empty sets respond most to rainfall, which the classifier uses heavily. Both signals should therefore be monitored together.

### 3.7 Agronomic recourse

Of the 2,772 recourse queries, only 62 (2.2%) were climate-compatible; in this dataset the climatic ranges of most crop pairs do not overlap. Recourse succeeded for 29.0% of compatible queries and 2.3% of incompatible ones (Figure 9b); the latter lie near the boundary of the compatibility test. Overall success was 2.9%. Successful switches cluster in a few pairs that differ mainly in soil (Figure 9a), such as coconut or orange to papaya, blackgram or lentil to grapes, kidneybeans to chickpea, and cotton to banana. The median indicative cost of a successful recourse was INR 13,914 per hectare (interquartile range 8,762–31,827).

Table 8 lists the six cheapest climate-compatible recourses. For a coconut field at 26.8 °C, 92.9% humidity and 225 mm rainfall, adding about 22 kg N, 12 kg P<sub>2</sub>O<sub>5</sub> and 17 kg K<sub>2</sub>O per hectare, with a 0.06-unit rise in pH, brought papaya into the conformal set at an estimated INR 3,482 per hectare. For a mothbean field, adding 16 kg N and 29 kg P<sub>2</sub>O<sub>5</sub> with a 0.09-unit rise in pH made lentil admissible at INR 4,034. The practical message is that most “why not my crop?” questions in this dataset have a climatic answer, which the system can now state explicitly, while a minority have a soil answer with a price attached.

### 3.8 Which measurements protect the guarantee

Figure 10(a) compares three permutation importances. By accuracy drop, humidity (0.297), nitrogen (0.224) and rainfall (0.184) lead, phosphorus and potassium follow at about 0.12, and temperature and pH barely register (below 0.011). The coverage-drop ranking differs. Humidity (0.630) stays first, but potassium (0.590) and phosphorus (0.562) move ahead of rainfall (0.541) and nitrogen (0.437), and temperature rises to 0.200. Shuffling temperature changes almost no top-ranked labels but removes a fifth of the coverage, because it lowers the forest's confidence enough to empty many sets. The crop-wise map in Figure 10(b) shows where each measurement matters: P and K carry apple and grapes, rainfall carries muskmelon and rice, and humidity carries chickpea and kidneybeans. For sensor procurement, these maps show which probes need the tightest error budget if the coverage guarantee is to hold in the field.

## 4. Discussion

The central finding is that on a benchmark where accuracy is close to its ceiling, the useful differences between methods lie elsewhere: in calibration, in coverage of the hardest crops and, above all, in behaviour under conditions that differ from the training data. A random forest with 99.5% accuracy produced conformal sets that missed jute one time in five, lost two thirds of its coverage under moderate sensor noise and did not react to a 6 °C warming. None of this is visible in a confusion matrix computed on a random split.

Each AgroConform component addresses one of these weaknesses at low cost. Split conformal prediction needs one extra pass over a calibration set. Adaptive sets (APS or RAPS) raised worst-crop coverage by about 12 percentage points relative to LAC, a statistically significant gain, at a cost of roughly 0.2 crops per set. Noise-aware calibration needs only an estimate of probe error, and overestimating that error is safer than underestimating it. The monitor's nearest-neighbour p-value is cheap to compute, and reporting ESS stops weighted sets from quietly becoming meaningless. Recourse turns an unwelcome recommendation into either a costed soil plan or a clear statement that the climate rules the crop out.

### 4.1 Practical implications and deployment guidance

The results translate into concrete guidance for the people who would build, run and use such a system; Table 9 summarises it by stakeholder. Three points deserve emphasis.

First, the advisory should show its uncertainty in a form a farmer can act on. A worked example from the test data illustrates this. One rice field (N 78, P 43, K 42 kg ha<sup>-1</sup>; 21.3 °C; 83% humidity; pH 7.3; 192 mm rainfall) received forest probabilities of 0.58 for rice and 0.42 for jute. At  $\alpha = 0.01$  the inclusion bar is a probability of about 0.42, so both crops just enter the set and the advisory reads “shortlist: rice or jute; choose by market price or rotation”. At  $\alpha = 0.05$  the bar rises to 0.68, neither crop clears it, and the advisory reads “no confident recommendation; please consult your extension officer”. The same field thus produces

a different but always honest message depending on how much risk the service is willing to automate. This matches the view of Hullman et al. [37] that prediction sets help decisions most when their meaning and the action they imply are stated plainly.

Second, the error level  $\alpha$  is a policy choice, not a technical detail. A service with few extension staff may accept  $\alpha = 0.05$  and refer about 5% of fields, while a service that wants near-complete automation can use  $\alpha = 0.01$ , referring about 1% of fields and offering occasional shortlists. In both cases the accuracy of automated recommendations stays above 99.6%.

Third, the sensor and calibration budget should follow the coverage-drop map rather than the accuracy ranking. Potassium and phosphorus probes, which look secondary by accuracy, carry much of the guarantee. Each probe type should be characterised once against laboratory analysis so that a realistic noise level can be used for noise-aware calibration. Where national soil-testing programmes such as India's Soil Health Card scheme already provide laboratory values, those records are a natural source of local calibration data. Calibration sets should hold at least 100 records per crop if per-crop guarantees at the 99% level are required.

#### 4.2 Limitations and future work

Several limitations should be stated plainly. The benchmark is balanced, regionally unspecified and partly curated, and its per-crop climate ranges are narrow. Real farm data will be noisier and less separable, which makes the conformal layer more, not less, important. Shifts and sensor errors were simulated rather than observed; the Gaussian noise model ignores bias, drift and correlated errors, which a field deployment should measure with paired laboratory analyses. Recourse costs are indicative unit prices and do not model crop response curves, soil buffering or the time amendments take to act, and nutrient removal is excluded. The coverage guarantee is an average over fields and does not hold for each individual farm; Mondrian or locally weighted variants narrow this gap only when enough local data exist. Natural next steps are validation on multi-season data from a specific agro-climatic zone, online updating of the conformal threshold as new seasons arrive [30, 31], group-conditional guarantees for soil and climate zones along the lines of Melki et al. [35], and linking the recourse engine to yield and price models so that switches are compared on expected profit rather than admissibility alone.

## 5. Conclusion

AgroConform treats crop recommendation as a question of reliability. Wrapping a standard classifier with split conformal prediction delivered coverage at the requested level and turned the output into a recommend, shortlist or abstain decision whose recommended crops were correct 99.6–99.98% of the time, depending on  $\alpha$ . Noise-aware calibration kept coverage at 95% across every tested level of soil-sensor error, where clean calibration fell as low as 38%. The shift monitor exposed climatic drift that the classifier ignored and showed when importance weighting had stopped being informative. Cost-aware recourse separated the few crop switches that soil amendment can achieve, each with a price per hectare, from the many that climate rules out. Coverage-drop importance identified the measurements on which the guarantee depends. Running in about 15 ms per field, the framework is practical for IoT-based advisory services and shifts the question from how accurate a model is to how far a specific recommendation can be trusted and what can be done about it.

#### Conflict of Interest

The author declares no conflict of interest.

#### Acknowledgments

The author thanks the anonymous reviewer and the editorial office for their evaluation of the manuscript. This research received no external funding.

#### Data Availability

The crop recommendation dataset used in this study is publicly available at <https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>. The analysis code and derived result files are available from the corresponding author on reasonable request.

## References

- [1] Elijah, O., Rahman, T. A., Orikumhi, I., Leow, C. Y., & Hindia, M. N. (2018). An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges. *IEEE Internet of Things Journal*, 5(5), 3758–3773. <https://doi.org/10.1109/JIOT.2018.2844296>
- [2] Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming – A review. *Agricultural Systems*, 153, 69–80. <https://doi.org/10.1016/j.agsy.2017.01.023>

- [3] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90. <https://doi.org/10.1016/j.compag.2018.02.016>
- [4] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674. <https://doi.org/10.3390/s18082674>
- [5] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE*, 11(6), e0156571. <https://doi.org/10.1371/journal.pone.0156571>
- [6] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, 10, 621. <https://doi.org/10.3389/fpls.2019.00621>
- [7] Sharafat, M. S., Kabya, N. D., Emu, R. I., Ahmed, M. U., Onik, J. C., Islam, M. A., & Khan, R. (2025). An IoT-enabled AI system for real-time crop prediction using soil and weather data in precision agriculture. *Smart Agricultural Technology*, 12, 101263. <https://doi.org/10.1016/j.atech.2025.101263>
- [8] Sulaeman, Y., Sutanto, E., Kasno, A., Sunandar, N., & Purwaningrahyu, R. D. (2024). Developing and testing a portable soil nutrient detector in irrigated and rainfed paddy soils from Java, Indonesia. *Computers*, 13(8), 209. <https://doi.org/10.3390/computers13080209>
- [9] Lobell, D. B., Schlenker, W., & Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science*, 333(6042), 616–620. <https://doi.org/10.1126/science.1204531>
- [10] Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, 114(35), 9326–9331. <https://doi.org/10.1073/pnas.1701762114>
- [11] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- [13] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
- [15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- [16] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [17] Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, 6, 257–265. <https://doi.org/10.1016/j.aiia.2022.11.003>
- [18] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. <https://doi.org/10.48550/arXiv.1706.04599>
- [19] Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2901–2907. <https://doi.org/10.1609/aaai.v29i1.9602>
- [20] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
- [21] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/22000000101>
- [22] Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic learning in a random world* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-031-06649-8>
- [23] Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>
- [24] Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002 (Lecture Notes in Computer Science, Vol. 2430, pp. 345–356)*. Springer. [https://doi.org/10.1007/3-540-36755-1\\_29](https://doi.org/10.1007/3-540-36755-1_29)

- [25] Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525), 223–234. <https://doi.org/10.1080/01621459.2017.1395341>
- [26] Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33. <https://doi.org/10.48550/arXiv.2006.02544>
- [27] Angelopoulos, A. N., Bates, S., Malik, J., & Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193* (published at ICLR 2021). <https://doi.org/10.48550/arXiv.2009.14193>
- [28] Vovk, V. (2013). Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2–3), 349–376. <https://doi.org/10.1007/s10994-013-5355-6>
- [29] Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1904.06019>
- [30] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 816–845. <https://doi.org/10.1214/23-AOS2276>
- [31] Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34. <https://doi.org/10.48550/arXiv.2106.00170>
- [32] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1), 521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>
- [33] Quiñero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2008). *Dataset shift in machine learning*. The MIT Press. <https://doi.org/10.7551/mitpress/9780262170055.001.0001>
- [34] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 227–244. [https://doi.org/10.1016/S0378-3758\(00\)00115-4](https://doi.org/10.1016/S0378-3758(00)00115-4)
- [35] Melki, P., Bombrun, L., Diallo, B., Dias, J., & da Costa, J.-P. (2023). Group-conditional conformal prediction via quantile regression calibration for crop and weed classification. *arXiv preprint arXiv:2308.15094*. <https://doi.org/10.48550/arXiv.2308.15094>
- [36] Melki, P., Bombrun, L., Diallo, B., Dias, J., & da Costa, J.-P. (2025). Uncertainty guarantees on automated precision weeding using conformal prediction. *arXiv preprint arXiv:2501.07185*. <https://doi.org/10.48550/arXiv.2501.07185>
- [37] Hullman, J., Wu, Y., Xie, D., Guo, Z., & Gelman, A. (2025). Conformal prediction and human decision making. *arXiv preprint arXiv:2503.11709*. <https://doi.org/10.48550/arXiv.2503.11709>
- [38] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>
- [39] Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10–19). <https://doi.org/10.1145/3287560.3287566>
- [40] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). <https://doi.org/10.1145/3351095.3372850>
- [41] Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2023). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5), 1–29. <https://doi.org/10.1145/3527848>
- [42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>
- [43] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81. <https://doi.org/10.48550/arXiv.1801.01489>

## Tables

**Table 1. Attributes of the crop recommendation dataset (n = 2,200; 22 crops × 100 records).**

| Attribute                | Min  | Max    | Mean  | SD    |
|--------------------------|------|--------|-------|-------|
| N (kg ha <sup>-1</sup> ) | 0.00 | 140.00 | 50.55 | 36.92 |
| P (kg ha <sup>-1</sup> ) | 5.00 | 145.00 | 53.36 | 32.99 |

| Attribute                | Min   | Max    | Mean   | SD    |
|--------------------------|-------|--------|--------|-------|
| K (kg ha <sup>-1</sup> ) | 5.00  | 205.00 | 48.15  | 50.65 |
| Temperature (°C)         | 8.83  | 43.68  | 25.62  | 5.06  |
| Relative humidity (%)    | 14.26 | 99.98  | 71.48  | 22.26 |
| Soil pH                  | 3.50  | 9.94   | 6.47   | 0.77  |
| Rainfall (mm)            | 20.21 | 298.56 | 103.46 | 54.96 |

**Table 2. Main symbols used in this article.**

| Symbol                 | Meaning                                                                                  |
|------------------------|------------------------------------------------------------------------------------------|
| $x, y$                 | Field attribute vector (N, P, K, temperature, humidity, pH, rainfall) and its crop label |
| $K$                    | Number of crops (22)                                                                     |
| $\hat{p}(k   x)$       | Probability that crop $k$ suits field $x$ , as estimated by the base learner             |
| $s(x, k)$              | Non-conformity score: how unusual crop $k$ would be for field $x$                        |
| $n$                    | Number of calibration records                                                            |
| $\alpha$               | Error level chosen by the user; $1 - \alpha$ is the target coverage                      |
| $\hat{q}$              | Conformal threshold computed from calibration scores (Eq. 1)                             |
| $C(x)$                 | Prediction set: the crops admitted for field $x$                                         |
| $\sigma, \hat{\sigma}$ | True and assumed soil-sensor noise level (in units of attribute SD)                      |
| $d(x), p(x)$           | Nearest-neighbour distance and its conformal p-value (Eq. 2)                             |
| $w(x), ESS$            | Covariate-shift weight and effective sample size of the weights (Eq. 3)                  |
| $\lambda$              | Intensity of the simulated covariate shift                                               |
| $c(x, x')$             | Indicative cost of moving field $x$ to amended field $x'$ (Eq. 4)                        |

**Table 3. AgroConform procedure.**

| Step | Operation                                                                                                                                                                                     |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Split the labelled data into training, calibration and test sets (50/25/25, stratified by crop).                                                                                              |
| 2    | Train the base learner $\hat{f}$ on the training set.                                                                                                                                         |
| 3    | Estimate probe noise $\hat{\sigma}$ for N, P, K and pH; if $\hat{\sigma} > 0$ , replace each calibration record by three noisy copies.                                                        |
| 4    | Compute calibration scores $s_i$ and the threshold $\hat{q}$ at the chosen $\alpha$ (Eq. 1).                                                                                                  |
| 5    | Fit the nearest-neighbour monitor and store calibration distances $d(x_i)$ .                                                                                                                  |
| 6    | For a new field $x$ : build $C(x)$ and compute $p(x)$ (Eq. 2).                                                                                                                                |
| 7    | If $ C(x)  = 1$ and $p(x) > 0.01$ , recommend; if $ C(x)  \geq 2$ , shortlist; otherwise abstain and refer.                                                                                   |
| 8    | If the farmer prefers crop $t \notin C(x)$ : search amended fields $x'$ with climate fixed and return the cheapest $x'$ with $t \in C(x')$ (Eq. 4), or report that the climate rules $t$ out. |
| 9    | When target-season fields are available, re-weight calibration scores, check ESS (Eq. 3) and recalibrate locally if ESS falls below about 100.                                                |

**Table 4. Five-fold cross-validated performance of the base learners (mean  $\pm$  standard deviation).**

| Model | Accuracy (%)     | Macro-F1 (%)     | NLL               | ECE (%)         | Brier             |
|-------|------------------|------------------|-------------------|-----------------|-------------------|
| LR    | 97.86 $\pm$ 0.47 | 97.85 $\pm$ 0.48 | 0.084 $\pm$ 0.008 | 3.68 $\pm$ 0.32 | 0.036 $\pm$ 0.005 |
| kNN   | 97.23 $\pm$ 0.72 | 97.21 $\pm$ 0.73 | 0.089 $\pm$ 0.021 | 2.47 $\pm$ 0.93 | 0.041 $\pm$ 0.009 |

| Model   | Accuracy (%) | Macro-F1 (%) | NLL           | ECE (%)      | Brier         |
|---------|--------------|--------------|---------------|--------------|---------------|
| SVM-RBF | 98.86 ± 0.14 | 98.86 ± 0.15 | 0.145 ± 0.001 | 11.06 ± 0.34 | 0.036 ± 0.001 |
| MLP     | 98.64 ± 0.29 | 98.63 ± 0.29 | 0.037 ± 0.008 | 0.99 ± 0.17  | 0.020 ± 0.003 |
| HistGB  | 98.95 ± 0.53 | 98.95 ± 0.53 | 0.037 ± 0.023 | 0.70 ± 0.31  | 0.014 ± 0.007 |
| RF      | 99.50 ± 0.30 | 99.50 ± 0.30 | 0.052 ± 0.002 | 4.13 ± 0.25  | 0.015 ± 0.001 |

**Table 5. Conformal prediction with the random forest, averaged over 30 random splits. “Worst crop” is the lowest class-conditional coverage among the 22 crops.**

| $\alpha$ | Method           | Coverage | Mean set size | Singleton (%) | Empty (%) | Worst crop |
|----------|------------------|----------|---------------|---------------|-----------|------------|
| 0.01     | LAC              | 0.993    | 1.00          | 99.4          | 0.3       | 0.921      |
| 0.01     | Mondrian-LAC     | 1.000    | 22.00*        | 0.0           | 0.0       | 1.000      |
| 0.01     | APS              | 0.991    | 1.82          | 58.0          | 0.8       | 0.943      |
| 0.01     | RAPS             | 0.991    | 1.55          | 58.0          | 0.8       | 0.943      |
| 0.01     | Naive cum. prob. | 1.000    | 2.80          | 32.0          | 0.0       | 1.000      |
| 0.05     | LAC              | 0.955    | 0.96          | 95.6          | 4.4       | 0.732      |
| 0.05     | Mondrian-LAC     | 0.964    | 0.97          | 96.0          | 3.3       | 0.816      |
| 0.05     | APS              | 0.950    | 1.16          | 79.4          | 4.9       | 0.848      |
| 0.05     | RAPS             | 0.950    | 1.14          | 79.4          | 4.9       | 0.848      |
| 0.05     | Naive cum. prob. | 1.000    | 1.59          | 64.6          | 0.0       | 1.000      |
| 0.1      | LAC              | 0.907    | 0.91          | 90.7          | 9.3       | 0.547      |
| 0.1      | Mondrian-LAC     | 0.930    | 0.94          | 93.1          | 6.7       | 0.764      |
| 0.1      | APS              | 0.895    | 0.98          | 82.2          | 10.3      | 0.751      |
| 0.1      | RAPS             | 0.895    | 0.98          | 82.2          | 10.3      | 0.751      |
| 0.1      | Naive cum. prob. | 1.000    | 1.27          | 79.8          | 0.0       | 1.000      |

\* With 25 calibration records per crop, the per-crop threshold at  $\alpha = 0.01$  is infinite and every set contains all 22 crops.

**Table 6. Conformal results at  $\alpha = 0.05$  with other base learners (10 splits each).**

| Base learner | Method       | Coverage | Mean set size | Worst crop |
|--------------|--------------|----------|---------------|------------|
| HistGB       | LAC          | 0.952    | 0.96          | 0.764      |
| HistGB       | Mondrian-LAC | 0.969    | 1.08          | 0.840      |
| HistGB       | APS          | 0.951    | 1.20          | 0.856      |
| SVM-RBF      | LAC          | 0.954    | 0.96          | 0.760      |
| SVM-RBF      | Mondrian-LAC | 0.956    | 0.99          | 0.788      |
| SVM-RBF      | APS          | 0.950    | 1.33          | 0.840      |
| kNN          | LAC          | 0.952    | 0.98          | 0.744      |
| kNN          | Mondrian-LAC | 0.972    | 2.35          | 0.852      |
| kNN          | APS          | 0.954    | 1.13          | 0.872      |

**Table 7. Three-way advisory outcomes (LAC with nearest-neighbour monitor at  $p \leq 0.01$ ), averaged over 20 splits.**

| $\alpha$ | Recommend (%) | Shortlist (%) | Abstain (%) | Accuracy of recommendations (%) |
|----------|---------------|---------------|-------------|---------------------------------|
| 0.01     | 98.6          | 0.3           | 1.1         | 99.62                           |

| $\alpha$ | Recommend (%) | Shortlist (%) | Abstain (%) | Accuracy of recommendations (%) |
|----------|---------------|---------------|-------------|---------------------------------|
| 0.02     | 97.8          | 0.0           | 2.2         | 99.75                           |
| 0.05     | 94.8          | 0.0           | 5.2         | 99.88                           |
| 0.1      | 90.1          | 0.0           | 9.9         | 99.98                           |

**Table 8. Cheapest climate-compatible recourses at  $\alpha = 0.05$  (adjustable attributes only; costs in INR ha<sup>-1</sup>, indicative).**

| From → to          | N (before → after) | P (before → after) | K (before → after) | pH (before → after) | Cost  |
|--------------------|--------------------|--------------------|--------------------|---------------------|-------|
| coconut → papaya   | 18 → 39.8          | 30 → 42.0          | 29 → 46.2          | 6.42 → 6.48         | 3,482 |
| mothbeans → lentil | 11 → 26.8          | 40 → 69.2          | 23 → 23.0          | 5.80 → 5.89         | 4,034 |
| orange → papaya    | 34 → 50.5          | 29 → 50.0          | 8 → 38.0           | 6.45 → 6.49         | 4,144 |
| orange → papaya    | 16 → 47.0          | 7 → 46.0           | 8 → 52.0           | 6.42 → 6.50         | 7,170 |
| coconut → papaya   | 39 → 45.0          | 29 → 41.8          | 29 → 46.2          | 6.14 → 6.41         | 8,503 |
| blackgram → maize  | 48 → 60.8          | 62 → 62.0          | 15 → 16.2          | 7.54 → 7.32         | 8,910 |

**Table 9. Deployment guidance derived from the results.**

| Stakeholder                   | Recommended practice                                                                                                                  | Supporting result                                                                   |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Farmer                        | Act on singletons; use local knowledge, prices or rotation to choose within a shortlist; treat abstention as a prompt to seek advice. | Recommended crops correct in $\geq 99.6\%$ of cases (Table 7)                       |
| Extension officer             | Review abstained fields first: they hold most of the model's errors.                                                                  | Top-ranked accuracy 84.4% on abstained vs 99.9% on recommended fields (Section 3.3) |
| Advisory service manager      | Set $\alpha$ by staffing capacity: $\alpha = 0.01$ refers about 1% of fields, $\alpha = 0.05$ about 5%.                               | Table 7                                                                             |
| Sensor procurement            | Prioritise accuracy of humidity, K and P measurements; characterise each probe against laboratory analysis.                           | Coverage-drop ranking (Figure 10)                                                   |
| Platform developer            | Calibrate with the expected probe noise; prefer APS or RAPS when every crop matters; log monitor flags and ESS.                       | Figures 5–7; Table 6                                                                |
| Researchers and policy bodies | Supply local, multi-season calibration data (e.g. from soil-testing programmes) with $\geq 100$ records per crop.                     | Mondrian sample-size rule (Section 3.2)                                             |

## Figures

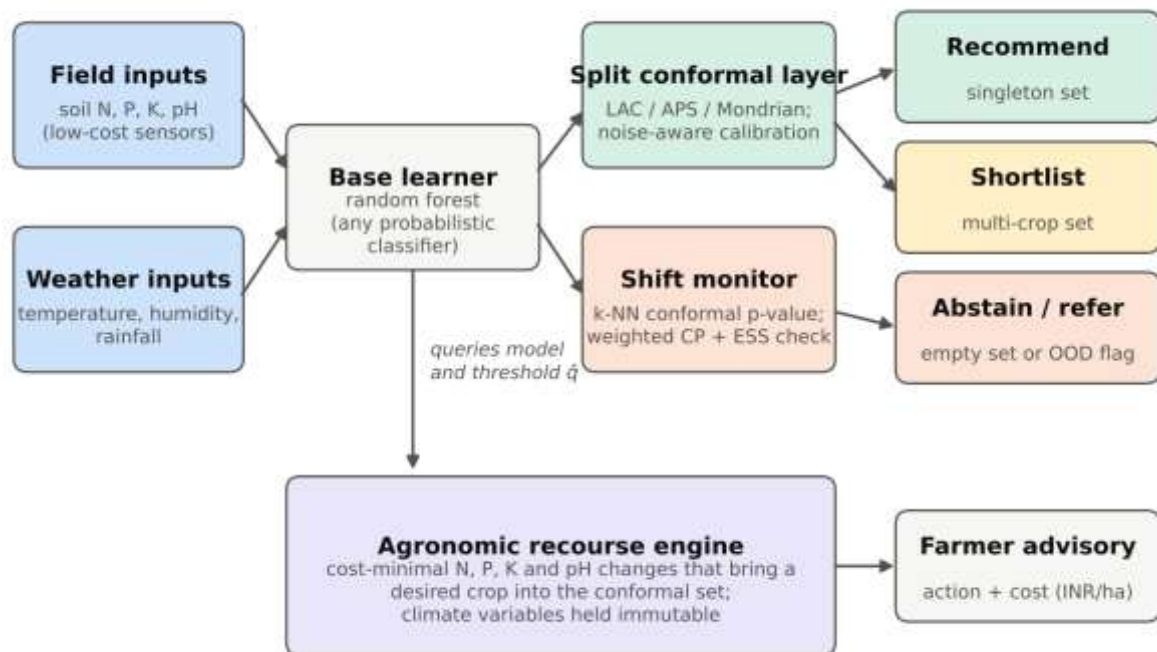


Figure 1. Architecture of the proposed AgroConform framework. A probabilistic base learner is wrapped by a split conformal layer and a shift monitor whose outputs drive a three-way advisory rule, and a recourse engine searches for cost-minimal soil amendments. Conformal components follow Angelopoulos and Bates [21] and Tibshirani et al. [29]; diagram prepared by the author.

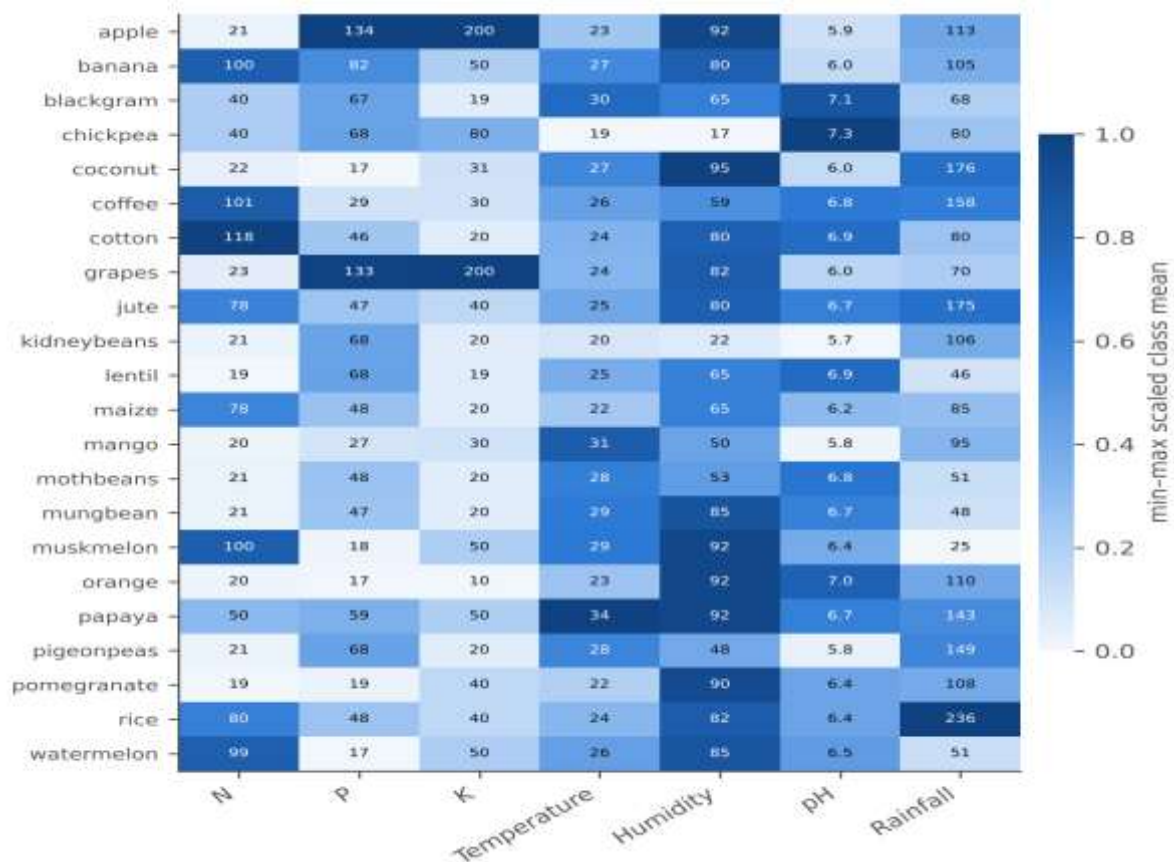


Figure 2. Per-crop mean of each attribute. Cell labels give the raw mean; colour encodes the mean after min-max scaling within each attribute so that columns are comparable. Data source: public crop recommendation dataset (Data Availability); figure prepared by the author.

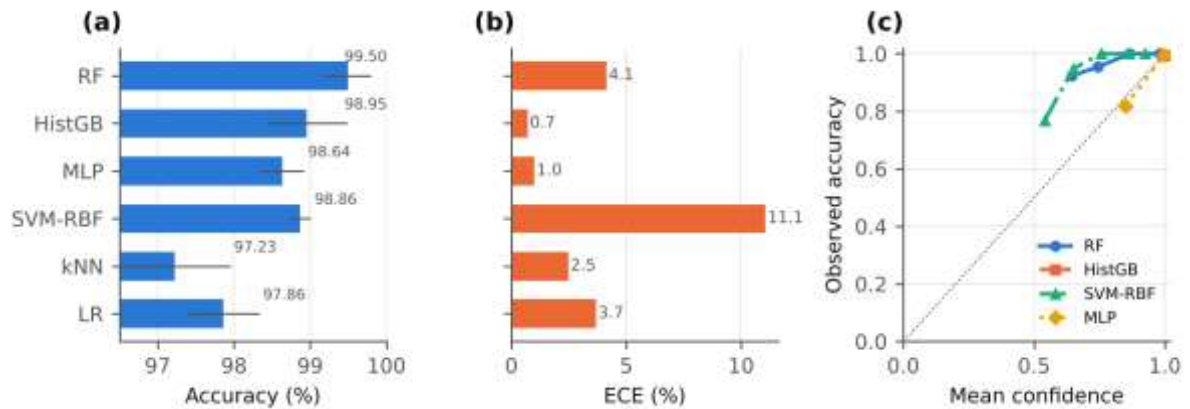


Figure 3. (a) Five-fold accuracy with standard deviation, (b) expected calibration error and (c) reliability diagram on a 30% hold-out for four learners. Points above the dotted diagonal indicate under-confidence. ECE follows Naeini et al. [19]; reliability analysis follows Guo et al. [18].

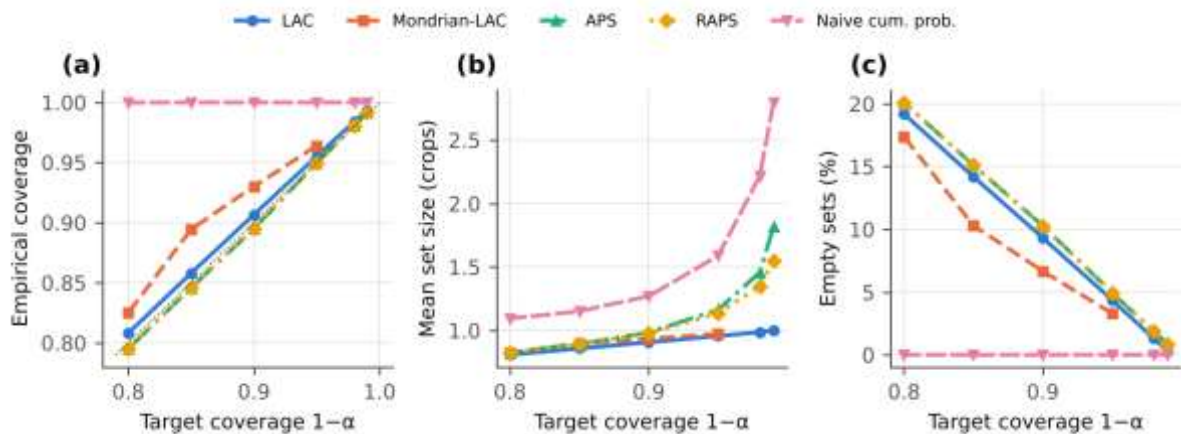


Figure 4. (a) Empirical versus target coverage, (b) mean prediction-set size and (c) share of empty sets for five set constructions with the random forest, averaged over 30 splits. Mondrian LAC is omitted at  $1 - \alpha \geq 0.98$  because its sets become trivial at that level. Methods follow Sadinle et al. [25], Romano et al. [26], Angelopoulos et al. [27] and Vovk [28].

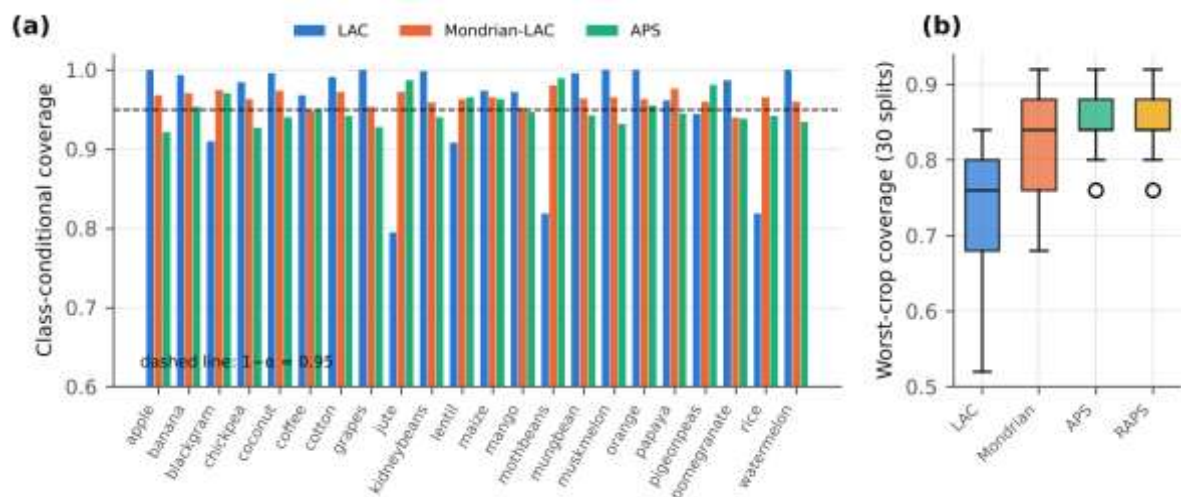


Figure 5. (a) Class-conditional coverage of each crop at  $\alpha = 0.05$ , averaged over 30 splits, and (b) distribution of the worst-crop coverage across splits. The dashed line marks the nominal 0.95 level. Mondrian conformal prediction follows Vovk [28].

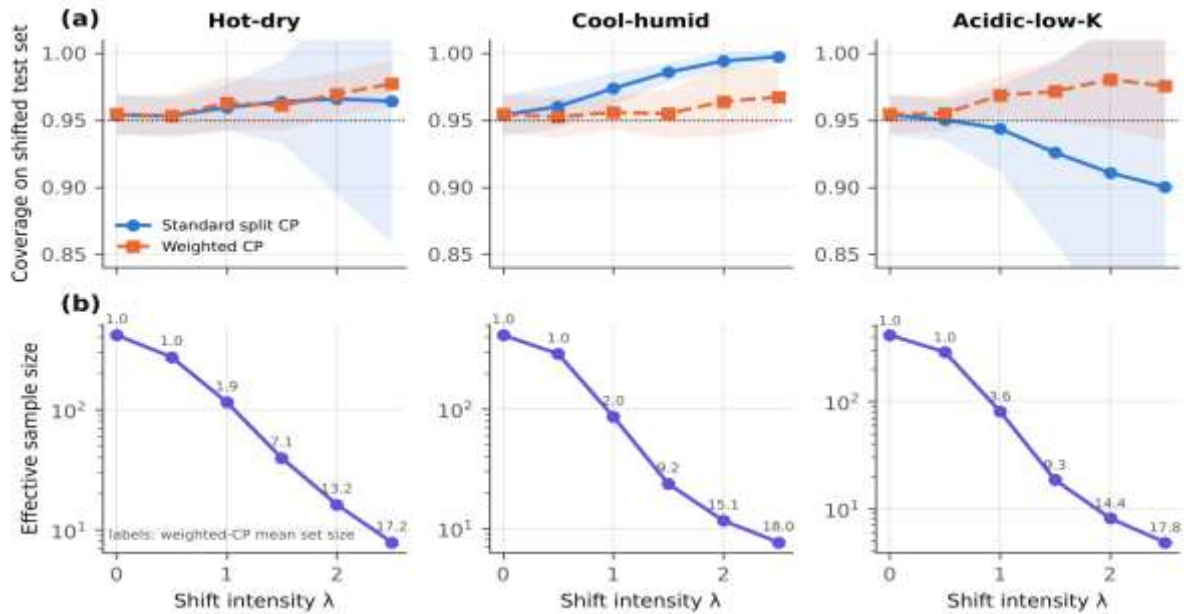


Figure 6. (a) Coverage of standard and weighted split conformal prediction under three simulated covariate shifts of increasing intensity  $\lambda$  (mean  $\pm$  SD over 15 repetitions; dotted line = 0.95). (b) Effective sample size of the calibration weights on a logarithmic axis; point labels give the mean weighted-set size. Weighted conformal prediction follows Tibshirani et al. [29].

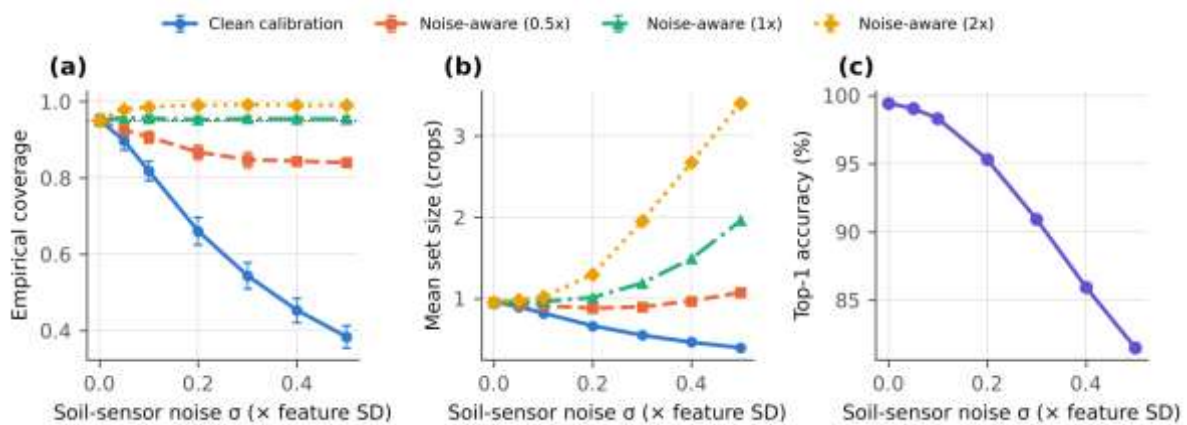


Figure 7. Effect of Gaussian soil-sensor noise on N, P, K and pH. (a) Coverage (mean  $\pm$  SD, 15 splits) of clean calibration and noise-aware calibration at half, equal and double the true noise level; (b) mean set size; (c) top-1 accuracy of the random forest. The dotted line marks  $1 - \alpha = 0.95$ .

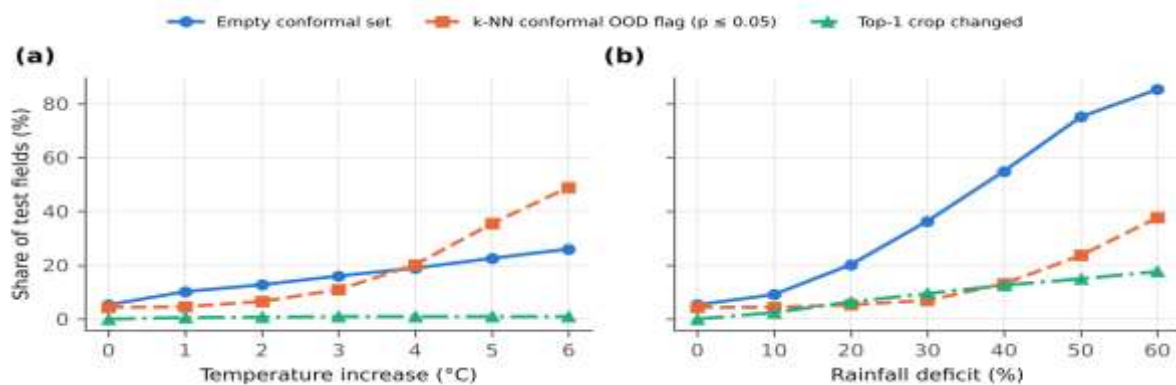


Figure 8. Response of the random forest and of the shift monitor to (a) uniform warming and (b) rainfall deficit applied to the test fields of one split. The conformal out-of-distribution flag uses the k-NN p-value of Eq. (2) at the 0.05 level.

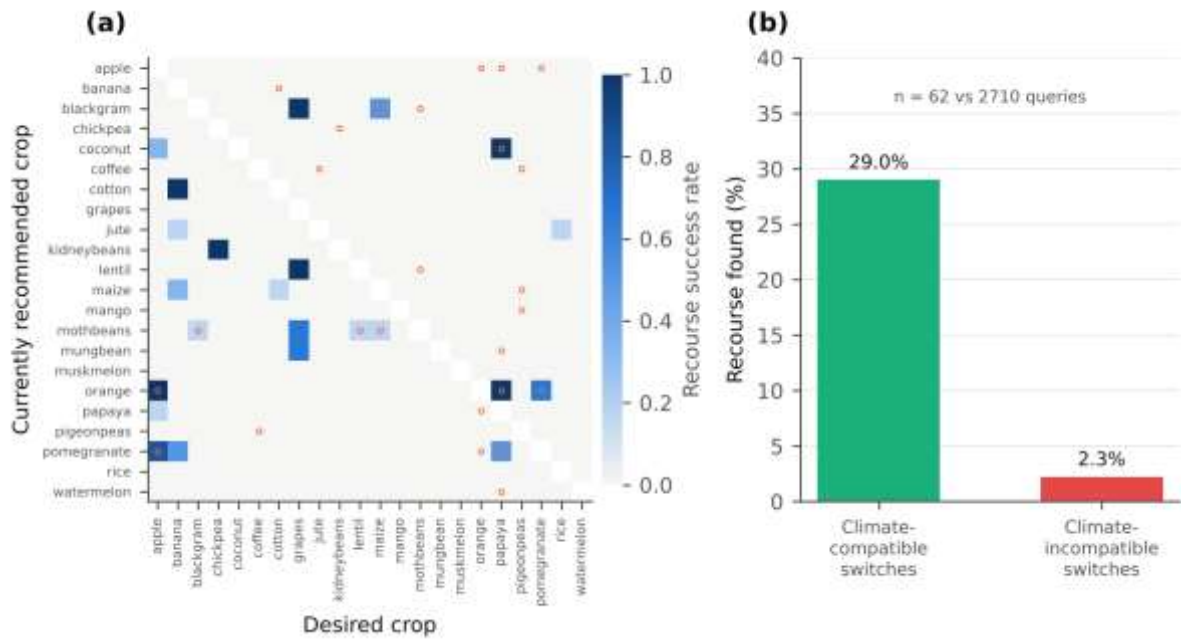


**Figure 9. (a)** Recourse success rate for every pair of currently recommended crop (rows) and desired crop (columns) at  $\alpha = 0.05$ ; orange rings mark pairs with at least one climate-compatible query. **(b)** Success rate for climate-compatible and climate-incompatible queries. Recourse formulation follows Ustun et al. [39] and Karimi et al. [41].

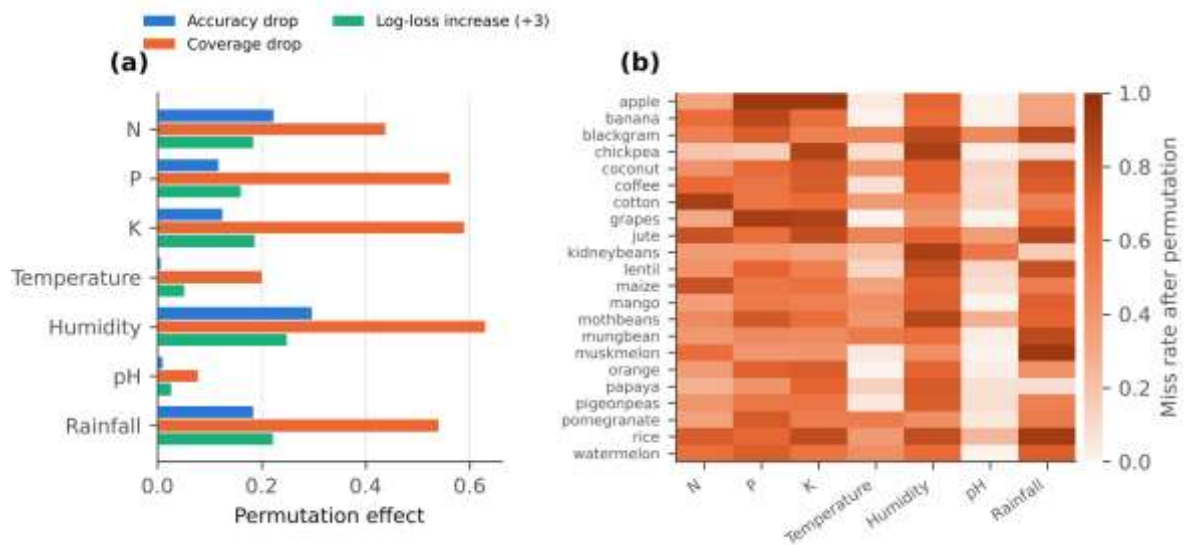


**Figure 10. (a)** Permutation importance measured as accuracy drop, coverage drop and log-loss increase (scaled by 1/3 for display), mean of 20 permutations. **(b)** Crop-wise miss rate of the conformal set after permuting each attribute. Permutation importance follows Fisher et al. [43]; coverage-drop importance is introduced in this work.