

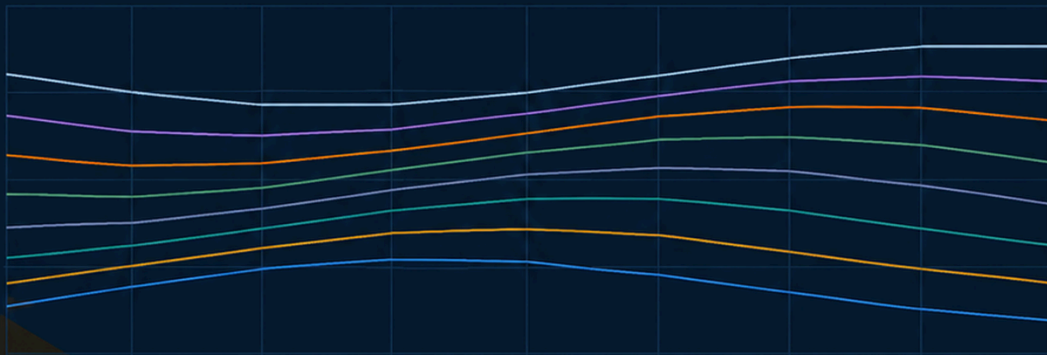


TALENCE
RESEARCH

RESEARCH METHODOLOGY REPORT

Development and Validation of a Food-Supply Shortage Early-Warning Model Across 43 African Countries

Integrated methodology for eight selected commodities, one-year-ahead risk estimation, historical analysis and country prioritisation



Prepared by Adewale Adelanwa
Published by Talence Research

talenceresearch.com

SEPTEMBER 2026

About this report

This report explains, as one continuous process, how the source data were narrowed, checked and transformed; how the shortage outcome and predictors were defined; how the model and warning threshold were selected and tested; and how the final method was applied to historical comparison and forward screening.

Two complementary report streams

This methodology report answers “How was the analysis built, tested and made defensible?” The separate analytical reports answer “What happened within each country and commodity, and which current cases require review?” Keeping these purposes separate prevents model-building detail from being confused with country conclusions.

Open research materials

Supporting analytical data and computational notebooks are published alongside this report for independent inspection and reproducibility. See Section 17.1 for access to the project repository.

Report structure

Part	Sections	What the reader will learn
I. Scope and data foundation	1-4	What the model predicts, which countries and commodities are covered, and how the analytical panel was constructed.
II. Outcome, predictors and development	5-10	How shortages were defined, how historical information became predictors, and how the model and warning threshold were selected.
III. Evaluation and forward application	11-15	How the locked model performed and how the same method produced evidence-qualified 2024 screening estimates.
IV. Ranking, safeguards and use	16-19	How countries were compared fairly, why rankings remain separate, and the limits governing interpretation.
Appendices	A-B	Plain-English definitions of technical terms and the primary methodological references.

Reading the numbers

Counts and percentages are reported at the precision needed to reproduce the analysis. Rounded prose descriptions are used only to make interpretation easier; the underlying tables retain the verified values.

Executive summary

The project uses FAO Food Balance Sheet information to estimate whether the calorie availability of one selected commodity in one country may fall materially in the following year. It does not attempt to predict famine, household hunger, food affordability or every dimension of food security. The target is deliberately narrower and measurable from the source data.

The workflow began with 373,670 raw FAOSTAT rows covering many countries, territories, aggregates and food categories. A controlled African scope was created using United Nations M49 statistical geography codes. The particular FAOSTAT release used contained usable Food Balance Sheet records for 43 African countries. Aggregate food groups were removed, population was separated from commodity rows, and eight important foods were selected using calorie relevance, country relevance and data coverage together.

The selected long dataset contained 77,994 source rows. It was converted into a complete country-commodity-year panel of 4,816 rows: 43 countries multiplied by 8 commodities multiplied by 14 years. Missing values, recorded zeros and negative accounting entries were retained as different data conditions. Missing values were handled through the model's preprocessing rules, while FAO quality flags were retained for data-quality assessment rather than used as direct predictors.

A shortage event was recorded only when three conditions were met: the starting-year supply was at least 5 kilocalories per person per day, the following year's supply fell by at least 15%, and the actual reduction was at least 5 kilocalories. Requiring both a percentage and an absolute fall prevented very small starting values or minor movements from being labelled as material shortages. This produced 334 events among 3,248 eligible annual transitions, an event rate of 10.28%.

The analytical panel was then converted into a model-ready feature table. It retained current supply, utilisation and nutrition measures; added previous-year values, year-to-year changes, and rolling three-year averages and variability where sufficient data were available; retained missing-value and source-availability indicators; and created supply-pressure and per-person measures. Columns with no variation in the training data were removed, leaving 129 predictors.

Logistic regression and random forest models were compared using time-ordered evidence. The random forest was selected because it showed the strongest average PR-AUC across earlier expanding-window checks. Its configuration and a 0.13 warning threshold were fixed before the final 2021-2022 predictor-year test was opened. On 498 untouched test cases, the model detected 33 of 51 shortages, missed 18 and issued 126 false warnings. Recall was 64.71%, precision 20.75%, ROC-AUC 0.7597 and PR-AUC 0.3071, almost three times the 10.24% no-skill reference.

After the unbiased evaluation was completed, an operational copy of the selected model was refitted using all 3,248 labelled historical observations. The exact 129-predictor structure was then reconstructed for all 344 country-commodity observations in 2023. Across 418,992 comparable historical feature values, the reconstruction produced no mismatches. Eligibility rules allowed 249

observations to be scored for possible shortages in 2024; 95 were deliberately left unscored because their 2023 calorie availability was missing or below the 5-kilocalorie minimum.

The unchanged operational method produced 77 formal warnings, including 12 high warnings. These are prospective screening estimates, not observed 2024 events. Evidence-support checks placed 195 scored rows within usual historical support, 53 under heightened review and one under limited evidence. The limited-evidence warning remained visible for manual review but was excluded from headline country ranking.

The analysis also created an observed historical event record and a method for comparing countries fairly. Historical burden, recent burden, average forward risk and highest forward risk remain separate rankings because each answers a different question. Historical rankings use conservative Wilson lower confidence bounds to account for unequal denominators, while forward priority uses evidence-qualified model probabilities. No composite country score was created.

Overall methodological conclusion

The process produced a credible but imperfect early-warning screening and reporting system. Its defensibility comes from traceable source selection, preserved data quality, leakage-safe timing, repeated out-of-time comparison, a threshold fixed before testing, one untouched final test, exact prospective-feature reconstruction, evidence-qualified ranking and honest reporting of both false warnings and missed shortages.

The complete methodology followed a controlled, auditable sequence

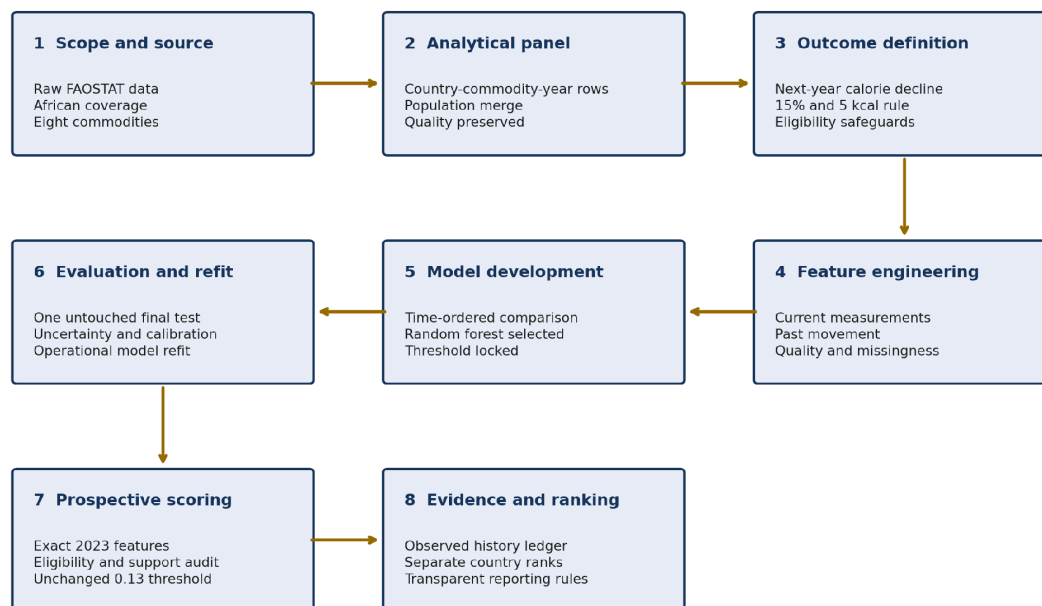


Figure 1. The methodology moved from raw source data to prospective scoring and analytical reporting through eight controlled stages.

1. Research purpose and boundaries

1.1 The question the model answers

For a country and one of eight selected commodities in year t , can information available in year t and earlier help identify a material fall in commodity-level food-supply calories per person in year $t + 1$?

The unit of analysis is therefore not simply a country. It is a country-commodity-year observation. A wheat observation for Ghana in 2021 and a maize observation for Ghana in 2021 are separate cases because the food systems, trade exposure and calorie contribution of the commodities differ.

1.2 Claims deliberately excluded

- The target is not a declaration of famine or a national food-security emergency.
- The data describe national apparent food availability, not what every household actually ate.
- A model warning is a request for investigation, not proof that a shortage will occur.
- The influential predictors are associations used for prediction; they are not automatically causes.

2. Source data and statistical meaning

The source was the flagged FAOSTAT Food Balance Sheet file covering 2010-2023. A Food Balance Sheet is a national accounting framework that brings together commodity supply and utilisation. Production, imports and stock movements contribute to supply; exports, feed, seed, processing, losses, non-food uses and food availability describe how that supply is used.

FAO describes Food Balance Sheets as national estimates of the pattern of food supply during a reference period. Per-person supply is obtained by relating the food quantity available for human consumption to the population and may be expressed in product weight, calories, protein and fat. This makes the source suitable for analysing national availability over time, but not for measuring differences between households, seasons or social groups.

Raw-data measure	Verified result
Rows	373,670
Columns	51
Areas	213
Item labels	120
Standard item codes	123
Elements	21
Years	14: 2010-2023
Flagged recorded values	E 54.32%; I 45.62%; X 0.06%

2.1 Why the quality flags matter

Flag	Plain-English meaning	Methodological treatment
E - Estimated	FAO produced an estimate rather than receiving a complete direct observation.	Retained and counted in row-quality information.
I - Imputed	A value was filled or derived where direct information was incomplete.	Retained and counted; not represented as directly observed.
X - External source	The value came from another source; population records carried this flag.	Retained and checked separately.

Important distinction

FAO imputation and modelling-stage imputation are different. An FAO “I” flag describes how the source value was produced. Modelling-stage imputation later replaces a missing predictor with a training-derived median so an algorithm can process the row. Both are recorded, but they should never be described as the same operation.

3. Geographic scope and commodity selection

3.1 Why 43 countries, not all African states

United Nations M49 numerical codes were used to define statistical geography. Codes were preferred to names because country names can vary across files, while the numerical code represents statistical identity. The African reference contained 54 sovereign states, but the FAOSTAT Food Balance Sheet release used in this project contained matching records for 43. Eleven expected countries were absent, and Western Sahara was treated as a separate territorial review case.

Geographic category	Count	Treatment
African sovereign states in reference	54	Reference universe
Countries represented in the FBS release	43	Included
Expected countries absent from the release	11	Documented source-coverage exclusions
Western Sahara	1	Absent and excluded as a separate territorial case

Correct wording for all later reports

The analysis covers the 43 African countries represented in the FAOSTAT Food Balance Sheet release used for this project. It must not be described as complete coverage of all African countries.

3.2 Removing aggregate categories

FAOSTAT contained both individual foods and aggregate categories such as Grand Total, Cereals excluding Beer and Starchy Roots. Including both would mix foods with groups that already contain those foods and could create double counting. Item-code evidence was therefore used to exclude 24 aggregate identities, leaving 97 individual commodity identities for selection.

3.3 How the eight foods were selected

Selection used several criteria together. Recent calorie importance was measured over 2021-2023, the latest three years available in the dataset, and weighted by population. Using a three-year average reduced the influence of an unusual result in one year while keeping the assessment relevant to recent food-supply patterns. The selection also considered the number of countries in which a food contributed at least 5 kilocalories per person per day, production and calorie-supply coverage, and whether the food served as a continent-wide or regional staple. No single continent-wide average controlled the choice.

Commodity	Relevant countries	Recent weighted kcal	Production coverage in relevant countries	Role
Maize and products	41	387.69	97.56%	Continental anchor
Wheat and products	43	364.51	62.46%	Continental anchor; import-dependent in many countries
Cassava and products	28	302.68	100.00%	Regional and continental staple
Rice and products	43	236.00	79.24%	Continental anchor
Sorghum and products	25	92.87	100.00%	Regional dryland cereal
Yams	13	83.14	97.80%	West African regional staple
Groundnuts	38	44.88	89.47%	Protein and oilseed staple
Millet and products	21	42.82	100.00%	Regional dryland cereal

The 5-kilocalorie relevance rule was used to decide whether a commodity was materially represented in a particular country. It was a practical selection and eligibility threshold, not a statement that values below 5 are nutritionally unimportant.

3.4 Changes in calorie availability over time

The selection table summarises each commodity's recent importance using the 2021-2023 average. Annual population-weighted calorie availability was also examined from 2010 to 2023 to show whether the recent position reflected a continuing pattern or a short-term movement.

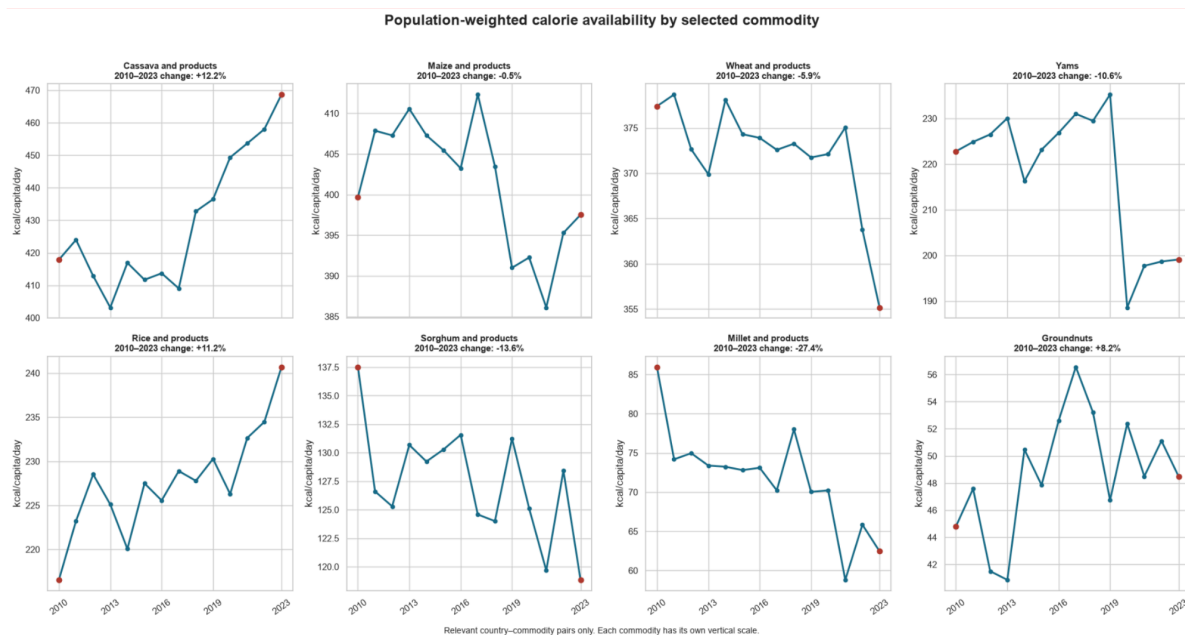


Figure 2. Population-weighted calorie availability for the eight selected commodities from 2010 to 2023. Each commodity uses its own vertical scale.

Cassava, rice and groundnuts increased overall; maize remained broadly stable; and wheat, yams, sorghum and millet declined. Millet recorded the largest overall reduction. These differences show why selection was not based on one year alone: the recent average described current importance, while the longer history showed how each commodity's contribution had changed.

4. Construction of the analytical panel

The selected long data were transformed into a panel in which every row represented one country, one commodity and one year. A complete grid was created before the source values were joined. This ensured that an absent source row remained visible as missing rather than disappearing from the analysis.

Panel component	Verified result	Why it matters
Countries	43	The covered geographic scope
Commodities	8	The controlled food set
Years	14	2010-2023
Expected rows	4,816	43 x 8 x 14 complete grid
Analytical elements	20	Supply, utilisation and nutrition measurements
Columns	84	Values, flags, source presence, population and quality fields
Duplicate final keys	0	Every country-commodity-year identity is unique

4.1 Population and unit reconciliation

Population was not treated as a food commodity. It was separated into a complete 602-row country-year table and merged by country and year. The source unit “1000 No” was converted to persons, and the relationship between the stored population fields was explicitly checked.

The project then recalculated the reported per-person food quantity, calorie, protein and fat measures from the national totals and population. The reconstructed results differed from the FAO-reported values by only 0.00% to 0.28% in typical comparisons. This close agreement confirmed that the population conversion, measurement units and country-year joins had been interpreted correctly; it was not used to replace the reported FAO values.

4.2 Availability and utilisation balance checks

Food Balance Sheets use accounting relationships to show where a commodity’s national supply came from and how that supply was used. The first relationship calculates the amount available within the country:

$$\text{Domestic supply} = \text{Production} + \text{Imports} - \text{Exports} - \text{Stock variation}$$

Production and imports increase available supply, while exports reduce it. Stock variation is subtracted because a positive value represents food moved into storage rather than made available for current use. Compared with FAO’s reported domestic-supply values, the median absolute difference was 0 thousand tonnes and 95% of comparable differences were no more than 1 thousand tonnes.

The second relationship checks whether domestic supply can be accounted for by its recorded uses:

$$\text{Domestic supply} = \text{Food} + \text{Feed} + \text{Seed} + \text{Processing} + \text{Other uses} + \text{Losses} + \text{Residuals} + \text{Tourist consumption}$$

This relationship also reconciled closely. Together, the checks confirmed the interpretation of units, positive and negative signs, and variable meanings. They were used only to validate the data structure, not to create or replace missing source values.

4.3 Why missing values were not automatically replaced with zero

A recorded zero and a missing value do not mean the same thing. A zero means that FAO explicitly reported the value as zero. A missing value means that no numerical value was available, but it does not by itself explain why. Each expected country-commodity-year value was therefore assigned to one of three states:

- Recorded: the source contained a numerical value. This includes positive values, recorded zeros and valid negative accounting values.
- Source-present missing: the relevant FAO source row existed, but the value for that particular year was blank.

- Source-absent: no source row existed for that country, commodity and measurement. The activity may not have applied, may not have been reported, or may have been unavailable; the data alone cannot distinguish these explanations.

Creating a complete grid meant that expected combinations remained visible even when no FAO row existed. Keeping the three states separate prevented unknown or unreported values from being silently converted into zeros.

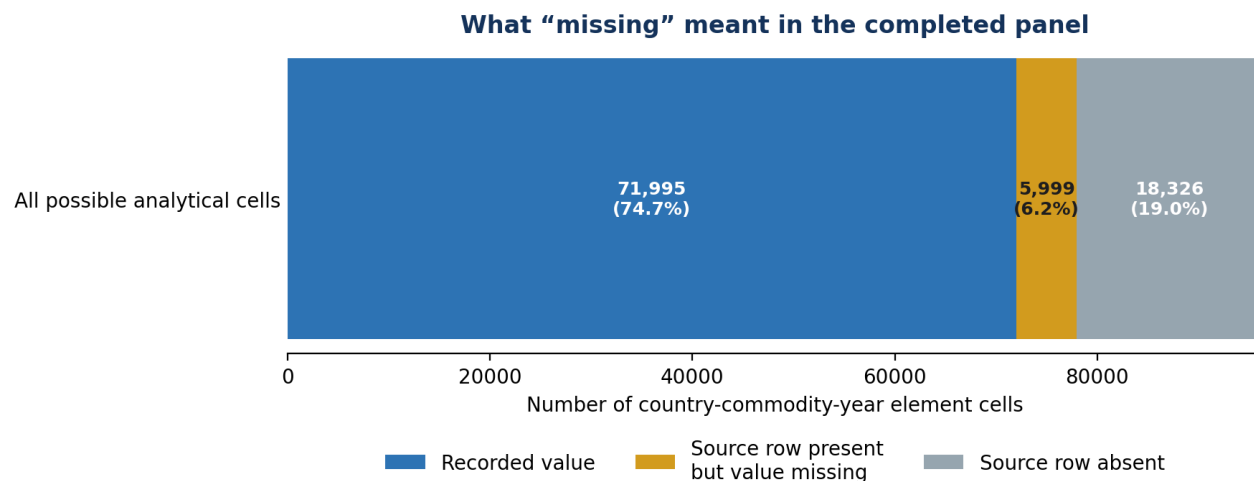


Figure 3. The panel preserved recorded values, source-present missing cells and source-absent cells as different data states.

Quality result	Count	Interpretation
Recorded analytical cells	71,995	Numeric values carried from the selected source data
Source-present missing cells	5,999	A source row existed but had no numeric value
Source-absent cells	18,326	No source row existed for that element and panel identity
Estimated cells	17,242	Recorded values carrying the FAO E flag
Imputed cells	54,753	Recorded values carrying the FAO I flag
Recorded zero cells	16,549	Real zeros, not blanks
Recorded negative cells	1,108	Mainly legitimate stock-variation and residual signs
Rows with all five principal outcomes	4,017	83.4% of panel rows

5. Definition of the prediction outcome

5.1 From annual movement to a labelled event

The model needed a clear answer to learn: shortage next year or no shortage next year. Each country-commodity calorie value was compared with the following year. Candidate rules combined a percentage decline with an absolute decline so that both proportional severity and real numerical size mattered.

Candidate definition	Events	Event rate	Interpretation
10% and 5 kcal decline	453	14.09%	Broad; includes more moderate movements
15% and 5 kcal decline	330	10.26%	Balanced provisional rule
20% and 5 kcal decline	244	7.59%	Stricter percentage rule
10% and 10 kcal decline	343	10.67%	Favours foods with higher starting calories
15% and 10 kcal decline	249	7.74%	Stronger absolute filter
20% and 10 kcal decline	184	5.72%	Severe and relatively rare

Final event definition

A positive event occurs when next-year food-supply calories per person per day fall by at least 15% and at least 5 kilocalories, provided the current-year value is at least 5 and the current and next-year values are both available.

The sensitivity table was produced during an earlier descriptive analysis restricted to country-commodity pairs considered recently relevant. Under that provisional filter, the selected rule identified 330 events among 3,215 eligible transitions. The same rule was later rebuilt without the future-derived relevance filter, producing the final leakage-safe total of 334 events among 3,248 eligible transitions. The definition did not change; only the eligible historical population became complete and time-safe.

5.2 Validation against related measures

The provisional events were checked against related analytical measures. All 330 had comparable per-person food quantity; 329 of those also declined, giving 99.70% agreement. Total food quantity declined in 98.48% of events and domestic supply declined in 91.82%. Production declined in 64.14%, showing that the target was not merely a production-failure label. Imports increased in only 22.26% of comparable events.

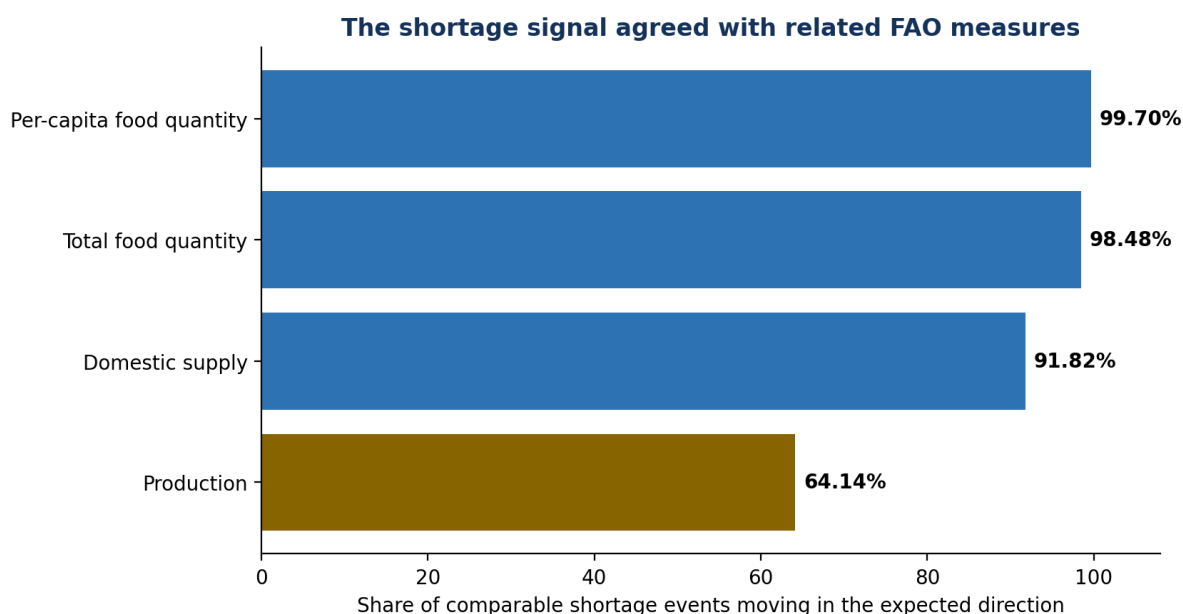


Figure 4. The provisional calorie-decline events agreed strongly with related food-availability measures.

5.3 Persistence after an event

Of the 330 provisional events, 304 could be followed for another year; the 26 events in 2023 had no later observation because the dataset ended. Among the 304 followed events, 36.84% declined further and 35.20% partially recovered but remained below 90% of the pre-event value. Only 18.42% fully recovered to at least the pre-event level. This supported treating many events as meaningful disruptions rather than harmless one-year numerical noise.

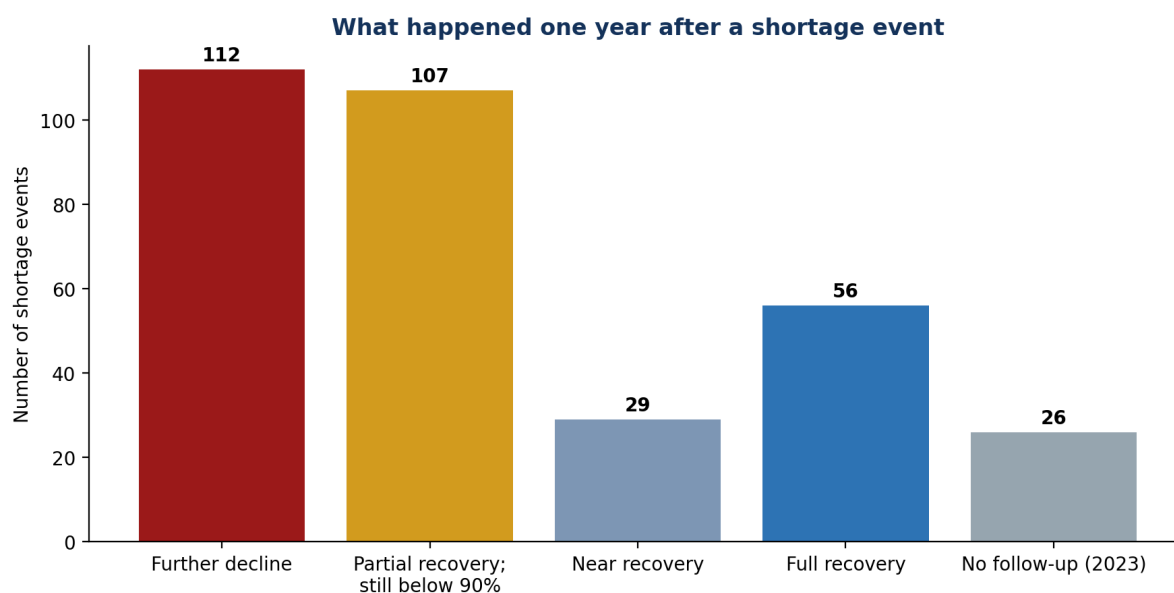


Figure 5. Most followed events remained materially below their pre-event level or declined further in the next year.

5.4 Leakage-safe target construction

The initial relevance label used average calorie availability from 2021-2023. Although useful for describing recent commodity importance, it would use future information if applied to earlier prediction years. It was therefore excluded from modelling eligibility. Eligibility was rebuilt separately for every year using only the current calorie value and the consecutive next-year value needed to calculate the outcome. The final supervised panel contained 3,248 eligible predictor rows and 334 events, with no missing targets and no ineligible row labelled as a non-event.

6. Time-ordered development design

A forecasting model should be tested in the same direction in which it will operate: using earlier information to estimate a later outcome. Randomly mixing years could allow later food-system patterns to influence a model being judged on earlier cases and make performance appear more favourable than it would be in practice. The data were therefore divided into chronological periods.

Because the model estimates risk one year ahead, every predictor year is paired with the following outcome year.

Period	Predictor years	Outcome years	Rows	Events	Purpose
Training	2010-2018	2011-2019	2,250	231	Fit candidate relationships
Validation	2019-2020	2020-2021	500	52	Compare models and thresholds
Locked test	2021-2022	2022-2023	498	51	One final unbiased evaluation

The training period allowed candidate models to learn historical relationships. The validation period compared model families, settings and warning thresholds on later observations. The locked test remained untouched until those decisions had been completed, providing one independent assessment on the most recent unseen historical cases.

Additional checks were conducted inside the training period using an established forecasting approach called expanding-window, or rolling-origin, validation. The amount of training history increased with each check while the immediately following predictor year remained unseen:

- Train on 2010-2014 and check predictor year 2015, whose outcome occurs in 2016.
- Train on 2010-2015 and check predictor year 2016, whose outcome occurs in 2017.
- Train on 2010-2016 and check predictor year 2017, whose outcome occurs in 2018.
- Train on 2010-2017 and check predictor year 2018, whose outcome occurs in 2019.

These repeated out-of-time checks showed whether performance was reasonably consistent and reduced dependence on one unusually easy or difficult period.

Meaning of “locked test”

The 2021-2022 outcomes were not used to choose the model family, tune its complexity or select the 0.13 warning threshold. The test was opened once after every decision had been recorded. Neither the model nor threshold was changed afterwards.

7. Feature engineering: converting the panel into model-ready predictors

Feature engineering converted understandable food-system measurements into a consistent set of predictors describing current conditions, previous values, recent changes, longer-term patterns, national context and data quality. Every predictor used information from the predictor year or earlier, ensuring that future knowledge could not enter the model.

7.1 Current-year analytical predictors

Seventeen current analytical measurements were retained: domestic supply, exports, fat per person, feed, food quantity, calorie supply per person, food quantity per person, imports, losses, other non-food uses, processing, production, protein per person, residuals, seed, stock variation and tourist consumption. Three national nutrition totals were excluded because they repeated information already represented by per-person measures and population.

7.2 Historical information

A current value shows conditions at one point in time, but it does not show whether the situation is improving, declining or fluctuating. Fourteen measurements recorded in at least 70% of the training data therefore received five historical summaries:

- Previous-year value: the measurement one year earlier.
- One-year absolute change: the current value minus the previous value.
- One-year percentage change: the change expressed relative to the absolute previous-year value.
- Three-year average: the average across the current and two preceding years, using at least two recorded values.
- Three-year variability: how much those same available values differed from one another.

Applying five summaries to fourteen measurements produced 70 historical features. Population also received a previous-year value and annual growth rate. Other non-food uses, processing and tourist consumption were retained only in their current form because their historical coverage was too sparse for reliable expansion.

7.3 Missingness and quality as information

Each retained analytical measurement received a missing-value indicator and a source-presence indicator. Six numeric row-quality counts and one categorical production-status feature were retained. Redundant or constant quality fields were excluded. This allowed the model to distinguish “a recorded zero” from “a value that is absent” and to recognise when a row relied more heavily on estimated or imputed source information.

7.4 Derived interpretable indicators

National totals can be difficult to compare because countries differ greatly in population and food-system size. Ten additional indicators therefore placed the quantities in context. Six expressed imports, production, exports, food use, losses and stock variation as percentages of domestic supply. For example, imports of 60 thousand tonnes against domestic supply of 100 thousand tonnes produce a 60% import-share indicator. These are accounting ratios, not slices of one pie, so they are not expected to total 100%.

Production, imports and exports were also converted to kilograms per person. The final indicator was log population, which compressed the very wide population range so that the largest countries did not dominate simply because their raw population numbers were much larger. The original population value remained available. Ratios were left missing when domestic supply was zero or unavailable, and no infinite values were permitted.

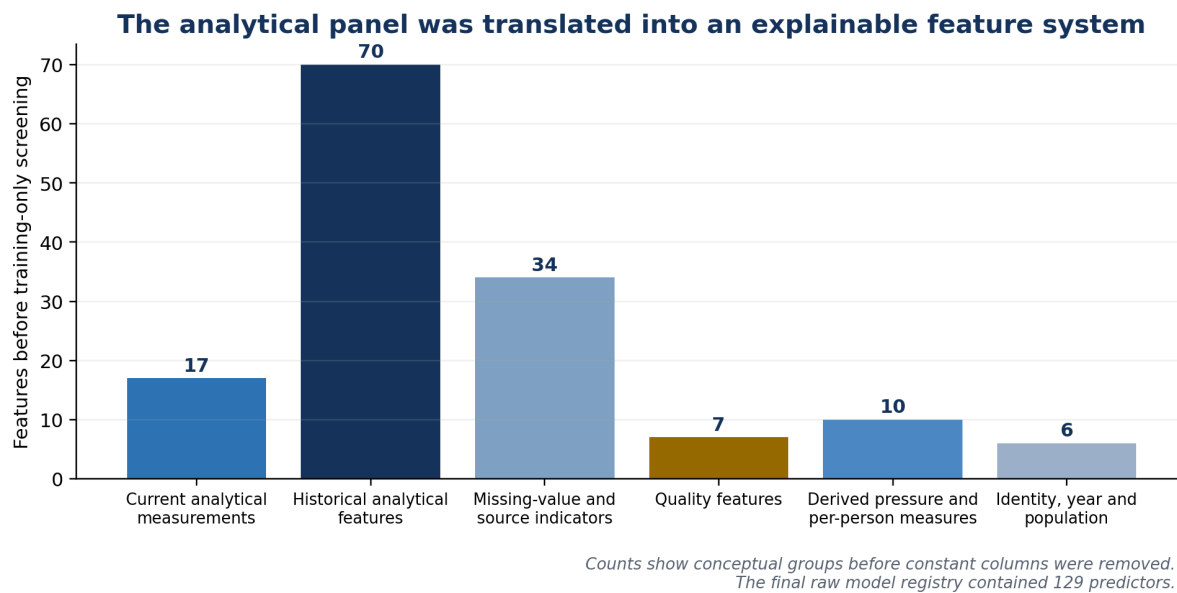


Figure 6. Distribution of the 144 candidate predictors across six feature families before training-only screening. The bar heights show feature counts, not predictive importance.

Feature-engineering checkpoint	Verified result
Rows	3,248
Columns before final modelling registry	152
Current analytical predictors	17
Historical analytical features	70
Derived current-year features	10
Candidate features screened	144
Constant training features removed	14
Final raw predictors	129
Categorical / binary / numeric	3 / 20 / 106
Duplicate modelling keys	0
Future information used	False

The six families contained 144 candidate predictors in total. Fourteen indicators showed no variation in the training period and one unstable percentage-change feature was excluded, leaving the final 129-predictor schema.

8. Training-only preprocessing

The 129 predictors were understandable to a human reader, but they were not yet in the numerical form required by the algorithms. Some were categories such as country and production status, some numerical values were missing, and the numerical scales differed greatly. Preprocessing converted them into a complete matrix that the models could process consistently.

Every preprocessing rule was learned from the 2010-2018 training period only and then applied unchanged to validation and test data. This prevented later periods from influencing missing-value replacements, category definitions, extreme-value boundaries or scaling.

- Categorical encoding: country, commodity code and production status were converted into separate 0/1 columns learned from the training categories.
- Missing numerical values: each blank was replaced with the training-period median for that predictor. For example, the training medians were 15 thousand tonnes for imports and 133 thousand tonnes for production. Existing missing-value indicators still showed the model that the original value had been absent.
- Extreme-value control: the ten derived ratios and per-person indicators were capped at their training-period 1st and 99th percentiles. Rows were not deleted; only unusually extreme numerical influence was limited.
- Scaling: numerical predictors were placed on comparable scales for algorithms affected by differences in numerical size. The scaling rules were calculated from training data and reused unchanged.

Processed matrix	Rows	Columns	Missing	Infinite
Training	2,250	181	0	0
Validation	500	181	0	0

The 181 processed columns are not a contradiction of the 129 raw predictors. The 126 non-categorical predictors remained individual columns, while country, commodity and production status expanded into 43, 8 and 4 indicator columns respectively. Therefore, $126 + 43 + 8 + 4 = 181$ columns in the numerical matrix seen by the algorithms.

9. Model comparison and selection

9.1 Why accuracy was not the main measure

Shortages represented about one observation in ten. A rule that always predicted “no shortage” achieved 89.6% validation accuracy while detecting none of the 52 shortages. Accuracy was therefore reported but not allowed to control model selection.

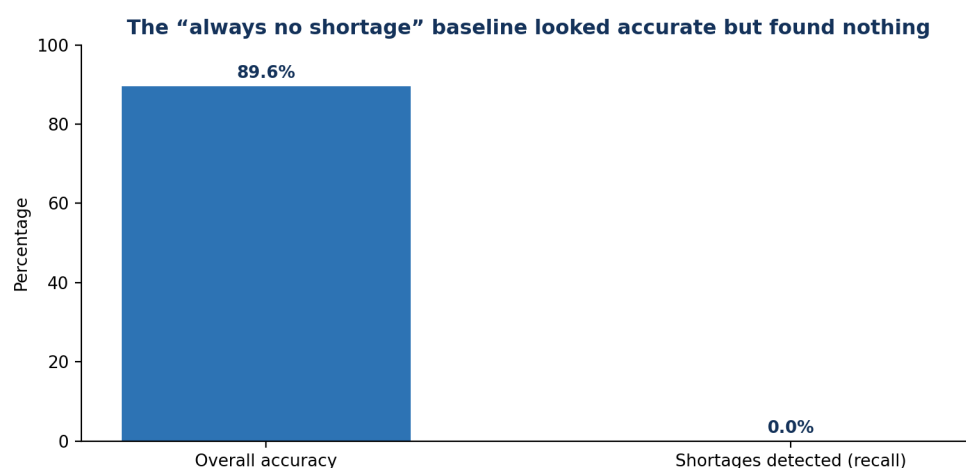


Figure 7. The no-shortage baseline appeared accurate only because shortages were uncommon; its recall was zero.

9.2 Measures used

Measure	Meaning in this project	Preferred direction
Recall	Share of real shortages that received a warning	Higher
Precision	Share of warnings that were followed by a shortage	Higher
F2	Combined score giving recall more importance than precision	Higher
Balanced accuracy	Average of shortage recall and non-shortage specificity	Higher
ROC-AUC	Ability to rank a shortage above a non-shortage across thresholds	Higher
PR-AUC	Precision-recall ranking quality for a rare outcome	Higher
Brier score	Average squared error in predicted probabilities	Lower

9.3 Logistic regression and random forest

Logistic regression provided a simpler benchmark. Random forest combined many decision trees and could represent nonlinear interactions among recent movement, supply, trade and commodity identity. Ordinary and class-balanced versions of both families were tested. Ordinary models treated every training row equally. Class-balanced models gave the less common shortage class more influence, using weights based on how frequently shortages and non-shortages appeared in the training data. The balanced option was tested rather than assumed to be better.

On the 2019-2020 validation period, the tuned logistic model produced a higher F2 and PR-AUC than the selected random forest. However, the earlier expanding-window checks told a broader story: the random forest achieved the highest mean temporal PR-AUC, 0.3317, compared with approximately 0.2612 for the selected logistic configuration. Temporal consistency was prioritised over winning on a single later slice.

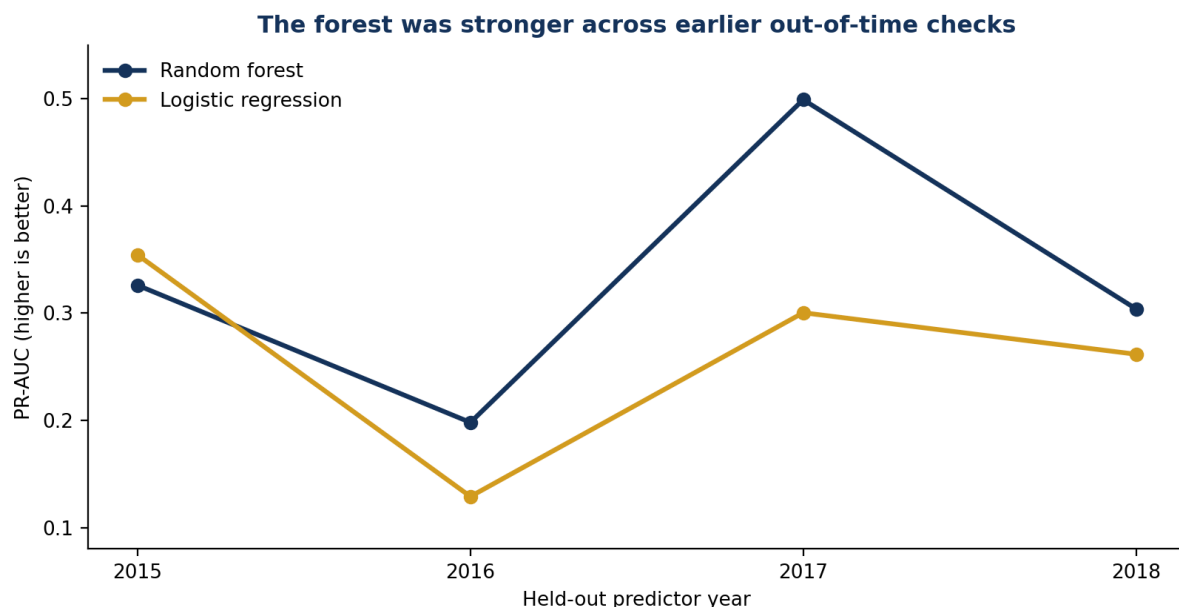


Figure 8. The random forest was stronger on average across the repeated earlier out-of-time checks, although performance varied by year.

Six random-forest complexity settings and seven logistic regularisation strengths were compared using the 2015-2018 temporal folds. The selected forest contained 500 trees, no explicit depth limit, a minimum of three observations in each final leaf, square-root feature selection at each split and no class weighting. An attempted probability ensemble assigned 100% of its weight to the forest, so it was rejected as unnecessary complexity.

10. Selecting the warning threshold

The random forest produces a probability rather than an automatic yes-or-no decision. A warning threshold converts that probability into action: at the selected threshold of 0.13, a country-commodity observation receives a formal warning when its estimated risk is 13% or higher.

A lower threshold detects more shortages but creates more false warnings. A higher threshold reduces the number of investigations and improves precision but misses more events. F2 was used to manage this trade-off because it combines recall and precision while giving recall four times as much emphasis in the calculation. This reflected the early-warning priority placed on avoiding missed shortages.

Thresholds were compared using 1,497 pooled out-of-time development predictions from predictor years 2015-2020, including 155 observed shortages. The maximum F2 occurred at 0.07, but its practical advantage was small relative to the additional warning burden:

Threshold	Warnings issued	Shortages detected	Recall	Precision	F2
0.07	788	126 of 155	81.29%	15.99%	0.4474
0.13	452	95 of 155	61.29%	21.02%	0.4431

Moving from 0.07 to 0.13 reduced warnings by 336, approximately 43%, while F2 fell by only 0.0043. The final rule therefore selected the most precise threshold among those achieving at least 99% of the maximum pooled development F2.

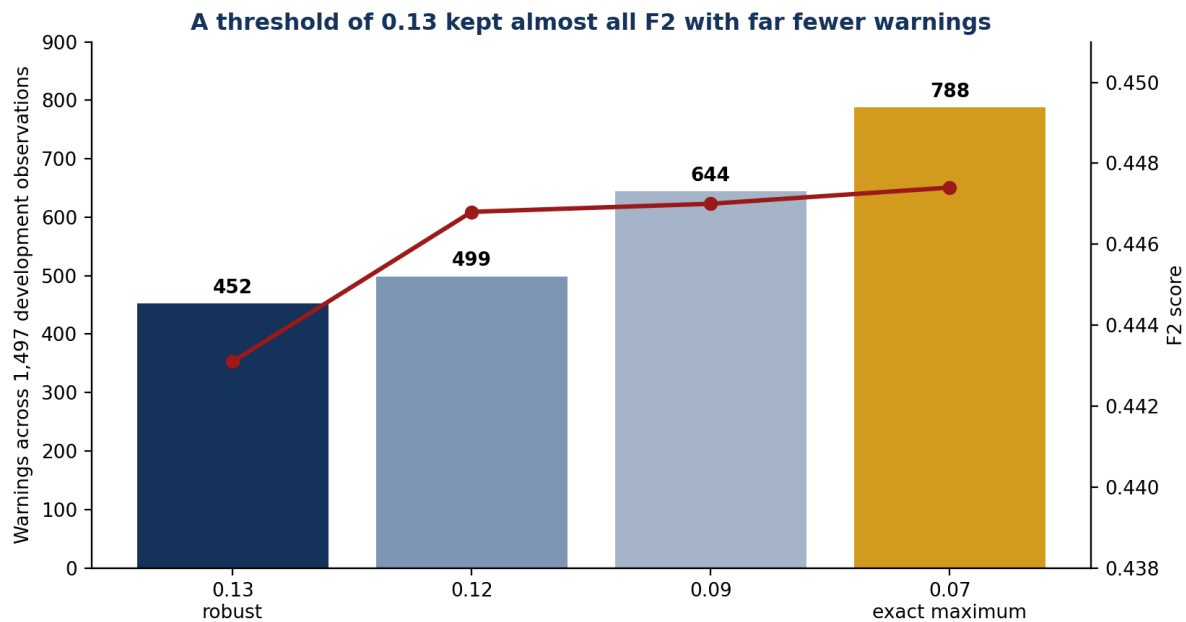


Figure 9. Warning volume increased sharply at lower thresholds, while the improvement in F2 was comparatively small.

In the chart, the bars use the left-hand axis and show how many warnings each threshold issued. The red line uses the right-hand axis and shows F2. Because the F2 axis covers a narrow range, the visual differences in the line appear larger than the underlying numerical changes.

Locked threshold rule

The final threshold was the most precise threshold among those achieving at least 99% of the maximum pooled development F2. It was locked at 0.13 before the final test was opened.

11. Final locked-test evidence

The final evaluation used predictor years 2021-2022 and outcome years 2022-2023. These 498 observations were used exactly once after the model and threshold were fixed.

Final-test outcome	Count	Plain-English meaning
Actual shortages	51	Cases meeting the pre-defined next-year decline rule
Warnings issued	159	Probabilities at or above 0.13
True positives	33	Shortages correctly warned about
False negatives	18	Shortages missed
False positives	126	Warnings not followed by the exact event
True negatives	321	Non-shortages correctly left without a warning

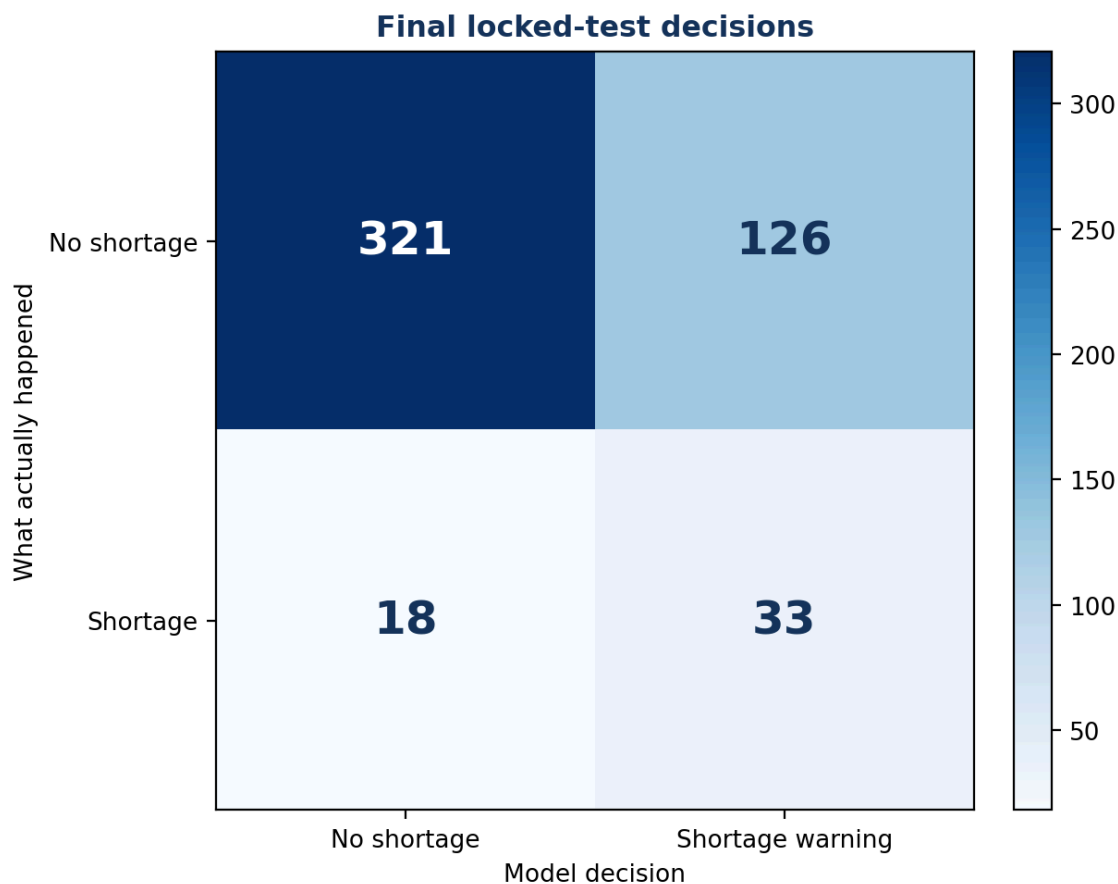


Figure 10. The final confusion matrix shows every correct and incorrect locked-test decision.

Measure	Estimate	95% confidence interval	Interpretation
Recall	0.6471	0.5000-0.7833	About 65% of shortages detected in this test
Precision	0.2075	0.1429-0.2761	About one warning in five was correct
Specificity	0.7181	0.6637-0.7699	About 72% of non-shortages received no warning
F2	0.4545	0.3431-0.5497	Recall-weighted warning performance
Balanced accuracy	0.6826	0.6075-0.7522	Average of recall and specificity
ROC-AUC	0.7597	0.6888-0.8241	Useful overall ranking
PR-AUC	0.3071	0.2083-0.4581	Almost 3x the 0.1024 no-skill reference
Brier score	0.0823	0.0645-0.1007	Useful but imperfect probability accuracy

The confidence intervals were produced by resampling complete country-commodity pairs rather than isolated rows. This respects the fact that observations from the same pair across two years may be related. The intervals describe sampling uncertainty; they are not guarantees of future performance.

11.1 Probability calibration and risk bands

The model assigned each test observation a probability, which was grouped into four operational bands: very low below 5%; watch from 5% to below 13%; warning from 13% to below 30%; and high warning at 30% or above. The 0.13 threshold separates monitoring from formal warning.

Calibration asks whether predicted probabilities resemble what actually happened. If a group assigned an average risk of about 20% experiences shortages about 20% of the time, that group is reasonably well calibrated.

Risk band	Test observations	Average predicted risk	Observed shortages	Observed shortage rate
Very low	162	2.94%	4	2.47%
Watch	177	8.46%	14	7.91%
Warning	118	19.02%	15	12.71%
High warning	41	43.60%	18	43.90%

Higher modelled risk generally corresponded to higher observed risk. Calibration was close in the very-low, watch and high-warning bands. The ordinary warning band overestimated risk: its average prediction was 19.02%, while 12.71% of observations became shortages. The probabilities are therefore useful for prioritisation but should not be treated as exact guarantees.

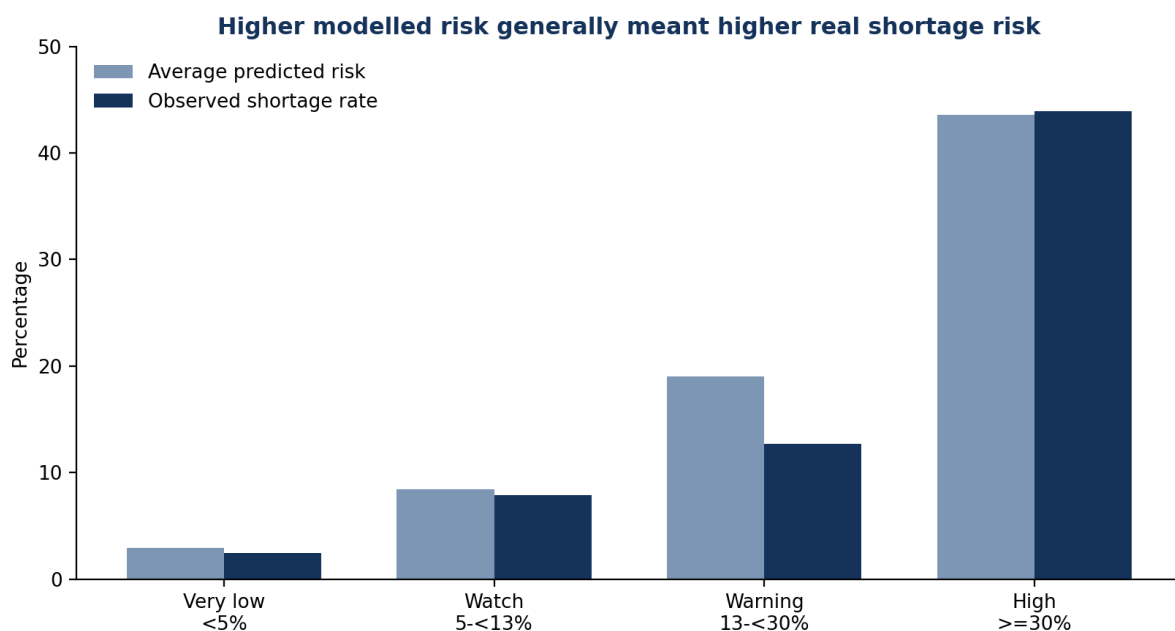


Figure 11. Average predicted risk and observed shortage rate within each locked-test risk band.

The light bar in each group shows average predicted risk; the dark bar shows the percentage that actually became shortages. Similar heights indicate closer calibration. Four of the 18 missed shortages were in the very-low band and 14 were in the watch band. Most misses therefore received some elevated concern but did not cross the formal threshold. This supports using the watch band for monitoring, but it does not justify changing the locked threshold after seeing the test results.

11.2 Subgroup cautions

Performance was stronger for maize, sorghum and groundnuts and weaker for wheat and cassava. Eight of the eighteen missed events were wheat shortages. However, several commodity test samples were very small: cassava had two shortages and yams had none. Commodity percentages

must therefore be read alongside event counts and should not be used to create new commodity-specific thresholds from the final test.

12. Understanding what influenced the model

Permutation importance examined which predictors the fitted model relied on during validation. One predictor was shuffled at a time, breaking its relationship with the outcome, and PR-AUC was recalculated. A larger fall meant that disrupting that predictor caused a greater loss in ranking performance.

The strongest information concerned recent percentage changes in total food quantity and calorie supply, changes in fat and protein supply, current and previous calorie levels, the three-year calorie average and commodity identity. Longer bars indicate greater predictive influence; they do not show whether a higher value increased or decreased risk.

The pattern suggests that recent direction, current conditions and short-term stability mattered more than a static total alone. Importance does not prove causation: a declining calorie measure may signal developing pressure without explaining what caused it.

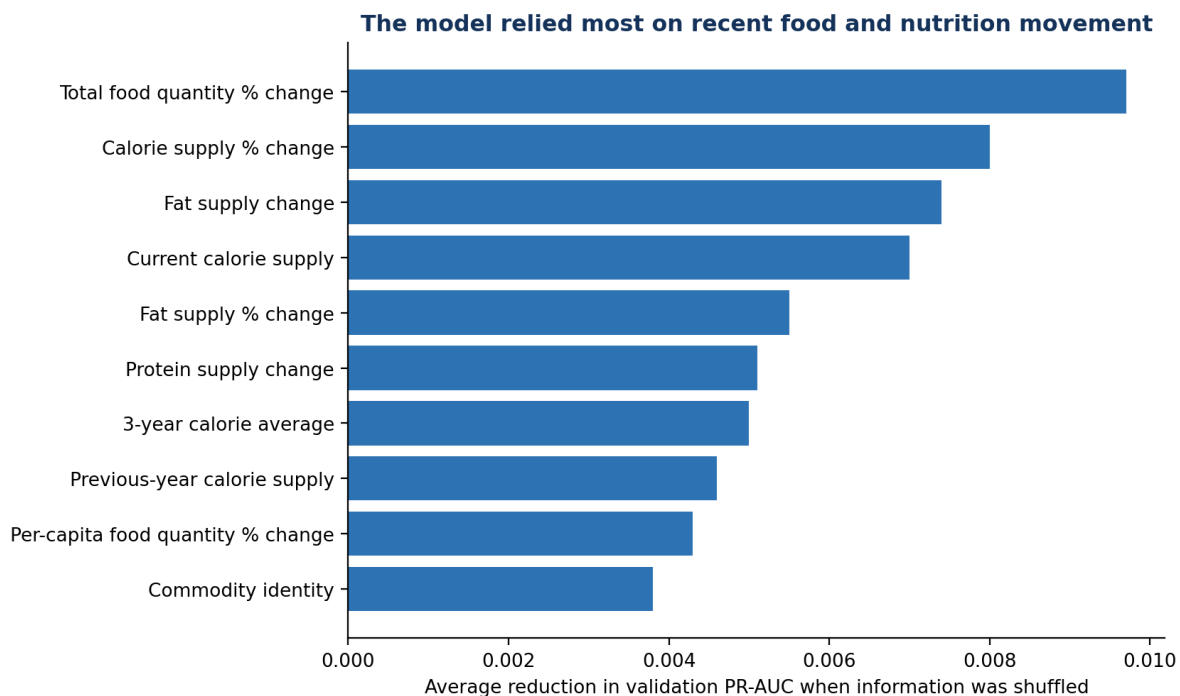


Figure 12. Reduction in validation PR-AUC after each predictor was shuffled. Longer bars indicate information on which the model relied more strongly.

13. Evaluation model and operational model

The evaluation model was fitted only through predictor year 2020 so that 2021-2022 could remain untouched. After the final evaluation was saved and declared complete, an operational copy was

fitted using all 3,248 labelled historical observations. This allowed the future-use model to learn from the most recent known examples without rewriting the official test result.

Operational-model record	Verified result
Training rows	3,248
Countries	43
Commodities	8
Country-commodity pairs	260
Predictor years	2010-2022
Outcome years	2011-2023
Positive events	334
Negative outcomes	2,914
Raw predictors	129
Probability threshold	0.13

Two models, two different jobs

The evaluation model answers “How well did the locked method perform on unseen later years?”

The operational model answers “Which fitted system should score the next genuinely future rows?” The operational copy must not be judged on 2021-2022 because those observations are now part of its training data. The saved locked-test result remains the unbiased evidence.

14. Reconstructing the 2023 predictors for forward estimation

The operational model was then connected to the latest available predictor year. The 2023 analytical panel contained the complete expected set of 344 observations: 43 countries multiplied by eight commodities, with no duplicate country-commodity-year records. Only 38 of the model's 129 required predictors were already present. The remaining 91, including previous-year values, annual changes, three-year histories, ratios and per-person measures, were reconstructed from the 2010-2023 analytical history.

Prospective feature checkpoint	Verified result
Available 2023 analytical rows	344
Countries / commodities	43 / 8
Locked model predictors	129
Predictors already present in the analytical panel	38
Predictors reconstructed from history	91
Historical feature cells compared	418,992
Historical mismatches	0
Infinite numeric values in 2023 features	0

The reconstruction rules were verified against the historical model-ready data rather than assumed from feature names. Percentage change used the absolute previous value as its denominator. Three-year means included the current year and required at least two recorded years. Three-year standard deviations used the same window, a minimum of two recorded years and the population

standard-deviation formula. Per-person quantities converted thousand tonnes to kilograms using a multiplier of one million, and log population used the natural logarithm of population plus one.

Why exact reconstruction matters

A fitted model is only meaningful when future rows use the same definitions, units, missing-value states, column order and data types used during training. Matching all 129 predictors across 418,992 comparable historical values provided direct evidence that the forward feature-building process reproduced the original method exactly.

14.1 Prospective scoring eligibility

Not every 2023 country-commodity row was eligible for a 2024 score. The same analytical relevance boundary used in target construction required current calorie availability to be recorded and at least 5 kilocalories per person per day. This rule protected the interpretation of a percentage decline and prevented absent evidence from being treated as low risk.

2023 scoring status	Rows	Countries	Commodities
Eligible and scored	249	43	8
Current calorie availability below 5 kcal	66	32	6
Current calorie availability missing	29	21	4

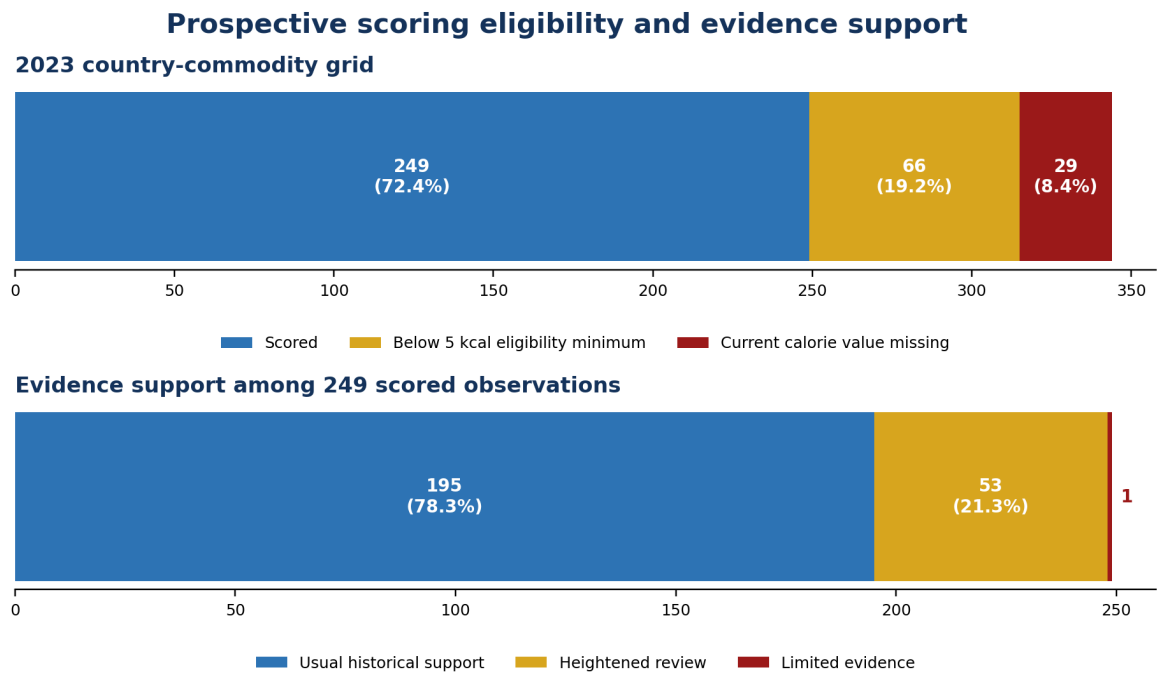


Figure 13. Prospective scoring eligibility was assessed separately from the strength of evidence supporting each score.

14.2 Applying the unchanged operational model

The 249 eligible rows were placed in the exact locked feature order and scored with the saved operational pipeline. The pipeline applied its saved preprocessing and fitted random forest. The

probability threshold remained 0.13; no 2024 score, country result or ranking was used to alter the model or threshold.

2024 risk band	Scored observations	Share of scored rows
Very low: below 5%	93	37.35%
Watch: 5% to below 13%	79	31.73%
Warning: 13% to below 30%	65	26.10%
High warning: 30% or more	12	4.82%

Prospective estimate, not observed outcome

The 77 formal warnings and 12 high warnings are model-generated screening estimates for target year 2024. No observed 2024 outcome was used. They must not be counted as historical shortage events or presented as proof that a shortage occurred.

15. Forward-score evidence and distribution audit

A probability can be technically valid while being supported by weaker-than-usual input evidence. Each scored row was therefore assessed using three questions: how much of the locked feature vector was missing, whether any non-calendar numerical feature lay outside the historical training range, and how much historical evidence existed for the same country-commodity pair.

Evidence-review level	Scored rows	Formal warnings	High warnings	Interpretation
Within usual historical support	195	59	9	Normal review process
Heightened evidence review	53	17	3	Score retained, but input evidence needs closer checking
Limited evidence: priority manual review	1	1	0	Insufficient support for headline ranking

The single limited-evidence case was cassava in Mauritius. It had no historical observations for the same country-commodity pair and three non-calendar features outside the training range. Its warning remained visible in the detailed register, but it was excluded from the headline country-ranking calculation. This preserved transparency without allowing a weakly supported row to influence a national summary.

The forward probability distribution was also compared with the untouched test distribution. Mean probability moved from 0.1206 in the locked test to 0.1088 in the forward register; the formal-warning share changed from 31.93% to 30.92%. The four-band Population Stability Index was 0.0316.

What the distribution check can and cannot show

The small descriptive shift indicates that the overall mix of risk bands did not change sharply. It does not measure predictive accuracy, prove calibration in 2024 or replace outcome-based monitoring when compatible 2024 data become available.

16. Historical evidence and country-ranking methodology

Known historical outcomes were kept separate from prospective predictions. The historical register contains 3,248 eligible country-commodity outcomes from 2011 to 2023, including 334 observed shortage events. The recent comparison uses the final five known outcome years, 2019-2023, containing 1,248 eligible observations and 137 events.

Ranking view	Evidence period	Primary ranking measure	Question answered
Historical burden	2011-2023	Lower Wilson boundary of observed event rate	Where has the most persistent evidence-supported burden been recorded?
Recent burden	2019-2023	Lower Wilson boundary of observed event rate	Where has burden been greatest in the final five known years?
Forward screening priority	2024 model register	Average evidence-qualified probability	Which country portfolios deserve priority review now?
Forward peak risk	2024 model register	Highest evidence-qualified commodity probability	Where is the strongest single commodity signal?

16.1 Why the Wilson lower boundary was used

Countries did not contribute identical numbers of eligible observations. Ranking raw percentages alone could place a country with a high rate based on very little evidence above a country with a similar rate supported by many more observations. The historical and recent burden ranks therefore order countries by the lower boundary of a 95% Wilson interval. This rewards high event rates that are also supported by a reasonable denominator.

A ranking aid, not a formal certainty statement

Country-commodity observations repeat across years and are not guaranteed to be statistically independent. The Wilson boundaries are therefore used as conservative denominator-aware ranking aids. They should not be interpreted as formal proof that one country's true long-run rate is higher than another's.

Countries with identical ranking values receive the same rank and the next position reflects the tie. Every table also retains the event count, eligible denominator, raw rate, interval boundaries, commodities assessed and most recent event year so the rank never appears without its evidence base.

16.2 How the forward ranks were constructed

For each country, the forward screening-priority rank uses the mean 2024 probability across scored commodities that passed the headline evidence-support rule. Averaging rather than summing prevents countries with more scored commodities from automatically receiving larger values. The number and share of formal warnings, high warnings and the maximum probability remain visible as supporting measures, not hidden ingredients in a composite score.

A separate peak-risk rank orders countries by their single highest evidence-qualified commodity probability. Average and peak risk are both needed: a country can have several moderate warnings

without one extreme score, or one exceptional commodity signal within an otherwise lower-risk portfolio.

16.3 Why no combined country score was created

Historical burden, recent burden, forward average risk and forward peak risk answer different questions and use different evidence. Combining them would require arbitrary weights and could make a model estimate look like an observed event. Their Spearman rank agreement confirms meaningful overlap but also substantial disagreement: historical and recent ranks agree at 0.796, while historical and forward-average ranks agree at 0.644.

The country rankings overlap, but they are not interchangeable

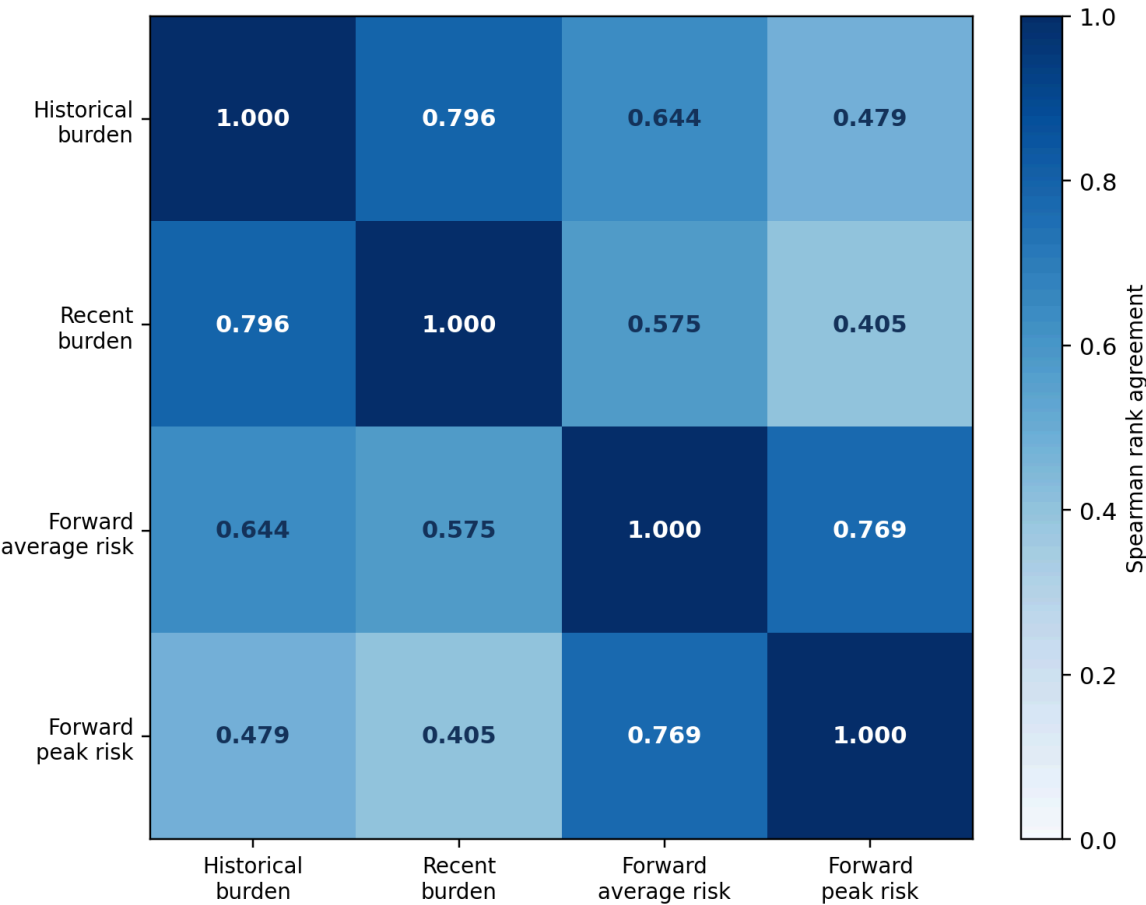


Figure 14. Separate country rankings overlap but remain materially different.

Reporting rule

The continental and country reports must display the historical, recent and forward rankings separately. A movement between ranks is an analytical finding to explain, not an error to average away.

16.4 Commodity comparisons

The same separation was applied to commodities. Historical and recent commodity burden use observed event evidence and denominator-aware boundaries. Forward commodity priority uses the average model probability across scored countries. This produced eight comparable commodity evidence records without treating a forward warning as a past event.

17. Methodological safeguards

Traceable source scope: The report states the exact FAOSTAT release period, raw dimensions, included countries, excluded geographic cases and eight commodity codes.

Identity controlled by codes: M49 and FAO item codes were used to resolve names, territories, aggregates and label-code collisions.

Complete panel before analysis: Expected country-commodity-year combinations were created first, so missing source combinations could not disappear.

Data quality retained: Source flags, recorded zeros, negative accounting values and distinct missingness states remained observable.

Outcome fixed before modelling: The 15% and 5-kilocalorie event rule was tested against alternatives and recorded before algorithm comparison.

Future information excluded: Targets, target years, split labels and next-year analytical values were excluded from predictors.

Time respected: Every important selection used earlier years to judge a later period; the final test remained untouched until the end.

Alternatives compared: A non-model baseline, logistic regression, random forest, class weighting, complexity settings and ensemble weights were compared.

Threshold locked: The 0.13 decision rule was selected from development predictions before test outcomes were inspected.

Prospective features reproduced exactly: All 129 future predictors matched the historical feature definitions across 418,992 comparable values before 2023 rows were scored.

Evidence support kept visible: Missingness, range checks and country-commodity history determined review levels; a limited-evidence warning remained visible but did not enter headline ranking.

Rankings kept conceptually separate: Observed historical burden, recent burden, forward average risk and forward peak risk were not collapsed into an arbitrary composite score.

Errors reported honestly: False warnings, missed shortages, confidence intervals, calibration limits and weak commodity groups were not hidden.

17.1 Open research materials and reproducibility

To support independent inspection and reproducibility, the supporting research materials for this project are published in the Early-Warning-Food-Prediction repository on GitHub. The repository contains the analytical data and computational notebooks used to develop, validate and apply the methodology described in this report.

The materials are intended to allow readers to examine the analytical workflow beyond the summary presented here, including source-data preparation, construction of the analytical panel, shortage-outcome definition, feature engineering, time-ordered model development, evaluation, prospective feature reconstruction and forward scoring.

The repository should be read alongside this methodology report. This report remains the primary explanation of the research scope, definitions, modelling decisions, safeguards, limitations and interpretation boundaries; the repository provides the underlying computational materials and implementation detail.

Research materials: [Early-Warning-Food-Prediction - data and analytical notebooks](#)

18. Limitations and appropriate use

National availability, not household consumption: Food Balance Sheets cannot show how food is distributed between households, regions, age groups or income groups.

Annual resolution: The system cannot detect a shock that begins and resolves within a year or provide monthly warning timing.

Publication delay: A one-year analytical horizon does not guarantee a full year of real-time operational lead because official data may be published later.

Estimated and imputed source data: The model inherits uncertainty from the underlying national accounts even though quality indicators are retained.

Narrow outcome: The target is a commodity-specific calorie-availability contraction, not a complete measure of access, affordability, nutrition or humanitarian severity.

False-warning burden: Only about one in five formal test warnings matched the exact target. Warnings should create a review queue, not trigger expensive action automatically.

Missed events: The model missed 18 of 51 test shortages. A no-warning decision means lower estimated risk, not safety.

Limited subgroup evidence: Country and commodity samples can be too small for stable local performance estimates.

Change over time: Climate, conflict, trade, prices and policy relationships can change; future performance must be monitored.

Prospective outcomes are unavailable: The 2024 register cannot yet show whether its warnings were correct because compatible observed 2024 outcomes were not used.

Ranking is purpose-specific: A country's position depends on whether the question concerns historical burden, recent burden, average forward risk or peak forward risk.

Intended use

The model is an early-warning screening and prioritisation tool. Its probabilities should be combined with weather, market prices, conflict, crop conditions, policy changes, trade disruption and local expert knowledge before any substantive conclusion is made.

19. Conclusion

The methodology forms one continuous chain from source selection to operational interpretation. It established a controlled 43-country and eight-commodity scope, preserved source quality and missingness, defined a measurable one-year-ahead outcome, created time-respecting predictors, compared alternative models, fixed the warning threshold before final testing, reproduced the feature system for 2023 and kept observed history separate from forward estimates.

The evidence supports a cautious operational conclusion: the model can rank and screen country-commodity cases for a material next-year contraction in calorie availability better than a no-skill reference, but it cannot replace a broader food-security assessment. Its value lies in disciplined prioritisation, transparent uncertainty and a clear distinction between observed events, prospective warnings and purpose-specific country rankings.

Appendix A. Glossary

Term	Meaning
Absolute change	The current value minus the previous value, expressed in the original unit.
Analytical panel	A table following the same country-commodity subjects through a sequence of years.
Balanced accuracy	The average of shortage recall and non-shortage specificity, preventing the common outcome from dominating the score.
Brier score	The average squared difference between predicted probabilities and actual 0/1 outcomes; lower is better.
Calibration	The degree to which predicted probabilities agree with observed event frequencies.
Categorical predictor	A predictor made of named groups rather than continuous numbers, such as country or commodity.
Class-balanced model	A model that gives the rarer shortage class more influence during training because shortages are less common.
Class imbalance	A situation in which one outcome is much more common; here, non-shortages greatly outnumber shortages.
Confidence interval	A range showing uncertainty around an estimate; it is not a guarantee of future performance.
Country-commodity-year	The basic observation: one commodity in one country during one predictor year.
Data leakage	Future or evaluation information entering model development when it would not exist at prediction time.
Domestic supply	The commodity quantity available nationally after accounting for production, trade and stock movement.
Evidence-review level	A label showing whether a forward score has usual historical support or requires closer manual checking.

Term	Meaning
Expanding-window validation	Repeatedly training on a growing block of earlier years and evaluating on the immediately following year.
F2 score	A combined precision-recall score that gives recall four times as much weight as precision in its formula.
False negative	A real shortage event that did not receive a formal warning.
False positive	A formal warning not followed by the exact shortage event defined in this project.
FAO imputation flag	An "I" flag showing that FAO filled or derived a source value; this differs from modelling-stage imputation.
Feature engineering	Creating model inputs from analytical measurements, such as changes, lags, averages and ratios.
Feature or predictor	A piece of information supplied to the model, such as current calorie availability or recent import change.
Formal warning	A model probability at or above the locked 0.13 threshold.
High warning	A model probability of 0.30 or higher.
Imputation	Replacing a missing modelling input with a value learned from training data, here the training median.
Lag	A value from an earlier year, such as the previous year's calorie availability.
Locked test	The final evaluation period kept untouched until the model and threshold had been selected.
Log population	Population transformed using the natural logarithm of population plus one to compress its numerical range.
Median	The middle value after observations are ordered; used here for missing-value replacement.
No-skill reference	The performance expected without useful ranking ability; for PR-AUC it equals the event rate.
Observed shortage event	A known historical outcome that met all parts of the predefined next-year calorie-decline rule.
One-hot encoding	Representing each category, such as maize or wheat, with a separate 0/1 column.
Operational model	The final copy refitted on all labelled historical observations for forward scoring.
Out-of-time evaluation	Testing on a later period than the data used for training.
Percentage change	The annual change divided by the absolute previous-year value and expressed as a percentage.
Permutation importance	The loss in model performance after one predictor is shuffled, used to assess predictive influence rather than causation.
Population Stability Index	A descriptive comparison of two score distributions; it is not a measure of predictive accuracy.
Population weighting	Giving countries influence in proportion to their populations when calculating a continental average.
Precision	The share of formal warnings that were followed by the defined shortage event.
PR-AUC	Precision-Recall Area Under the Curve: ranking quality focused on the rare positive outcome; higher is better.
Probability threshold	The probability at or above which the model issues a formal warning.
Prospective warning	A model-generated estimate for a future target year; it is not an observed event.
Random forest	A model that combines predictions from many varied decision trees.
Recall	The share of all real shortage events detected by the model.
Risk band	A probability range used to translate model scores into very-low, watch, warning or high-warning categories.
ROC-AUC	The probability that a randomly selected shortage receives a higher score than a randomly selected non-shortage.
Rolling three-year average	The average of the current and two preceding years, requiring at least two recorded values.
Rolling three-year variability	The population standard deviation of available values within the current and two preceding years.
Source-absent	No FAO row existed for the expected country, commodity and measurement combination.
Source-present missing	The FAO row existed, but its value for the relevant year was blank.
Specificity	The share of non-shortages correctly left without a warning.
Target or label	The outcome the model learns to predict; here, the predefined next-year calorie-availability contraction.
Temporal split	Dividing data by time so that later years evaluate models trained on earlier years.
True negative	A non-shortage correctly left without a formal warning.
True positive	A shortage that correctly received a formal warning.
Watch band	A probability from 5% to below the locked 13% formal-warning threshold.
Wilson lower boundary	A conservative rate boundary used here to reduce the ranking advantage of small denominators.

Appendix B. Primary methodological references

Food and Agriculture Organization of the United Nations. FAOSTAT data portal. <https://www.fao.org/faostat/en/>

Food and Agriculture Organization of the United Nations. Food balance sheets and supply utilization accounts resource handbook. <https://openknowledge.fao.org/handle/20.500.14283/cd4472en>

Food and Agriculture Organization of the United Nations. Food Balance Sheets: A Handbook. <https://openknowledge.fao.org/handle/20.500.14283/x9892e>

United Nations Statistics Division. Standard country or area codes for statistical use (M49). <https://unstats.un.org/unsd/methodology/m49/>

Project research materials

Adelanwa, A. (2026). *Early-Warning-Food-Prediction: Supporting data and analytical notebooks for the Food-Supply Shortage Early-Warning project*. GitHub.

Repository: [Early-Warning-Food-Prediction](#)

All numerical project results were checked against the validated analytical and model outputs used to construct this report. External references support the interpretation of Food Balance Sheets and M49 statistical geography; they do not replace the project's computational evidence. Supporting analytical data and computational notebooks are published separately through the project's GitHub repository to provide an inspectable record of the analytical workflow.