HSAT

# Ensuring Accurate Models

## Summary

HSAT's yield forecasting approach combines diverse data inputs—including satellite imagery, weather data, economic indicators, market prices, and policy information—to model the multiple factors influencing agricultural output. These data are used to train machine learning models that identify spatial and temporal patterns relevant to crop performance.

To assess performance when developing model formulation, we apply leave-one-out cross-validation, ensuring that each historical prediction is tested only on data that would have been available at the time. This avoids data leakage and supports a realistic evaluation of predictive accuracy.

Forecasts are issued on a weekly basis and stored under a strict point-in-time protocol. Each output is time-stamped and preserved without modification, allowing for full historical reconstruction and auditability.

This structure supports back testing and analysis workflows where temporal consistency is essential.

The underlying dataset covers a wide range of crops and geographies. For example, for maize alone, HSAT maintains over **943,000 weather-observation days** and **630 million hectares of satellite coverage.**

HSAT maintains similarly extensive datasets for wheat, rice, soy, sugar, cocoa, coffee, palm, canola, and sunflower, with up to **18 countries** per crop. Additional crops, including asparagus, aubergines, cotton, grapes, and tomatoes, are also tracked with varying levels of spatial and temporal coverage, available on request.

This infrastructure supports high-frequency, forward-looking yield estimates across key global regions, grounded in both agronomic signals and real-world economic context.

# Contents

Summary ..... 0
Overview of Our Process ..... 2
Validation – Real Data, No Data Leaking, Point in Time ..... 2
Preventing Data Leakage ..... 3
Real Data ..... 3
Point in Time ..... 3
Appendix 1: Data. Delivery. File Formats ..... 4
Appendix 2: Crop Coverage ..... 5
Grains and Cereals ..... 5
Oilseeds ..... 5
Soft Commodities ..... 5
Nuts ..... 5
Fruit Crops ..... 5
Root and Tuber Crops ..... 5
Legumes and Pulses ..... 5
Horticultural and Specialty Crops ..... 5

## Overview of Our Process

HSAT's yield models combine a wide range of data sources to capture the complex drivers of agricultural production. These include:

- **Satellite imagery** to monitor crop development, vegetation health, and land use.
- **Weather data** such as rainfall, temperature, and drought conditions.
- **Macroeconomic indicators** and **local economic data** to account for market context and regional pressures.
- **Commodity prices**, **foreign exchange rates**, and **input costs** (e.g. seeds, fertiliser, fuel) to reflect affordability and investment in agriculture.
- **Government policies** and subsidy regimes that influence planting decisions and production incentives.

We use these inputs in machine learning models trained to detect patterns across space and time. The models are validated using leave-one-out cross-validation to simulate real-world forecasting conditions and are deployed under a strict point-in-time process—ensuring that each prediction is made only with the data available at that moment and preserved without revision.

This approach allows us to generate forward-looking, high-frequency yield forecasts that respond to both environmental signals and economic realities.

### Validation – Real Data, No Data Leaking, Point in Time

When testing we ensure that our models are accurate reflections of the real world, our key principles are:

- Real world data
- No data leakage
- Point in time

This means that when we run a test to predict a year, we have not seen that data before and we are simulating real world events, with only the data available at that time.

**Example**: When testing a model by simulating a June 2008 run to predict the December 2008 harvest, we exclude all 2008 harvest observations and use only the inputs (weather, economic, satellite, etc.) that were available up to June 2008.

# HSAT Predictive Models

## Preventing Data Leakage

To prevent data leakage—where the model remembers the data it was trained on—whilst maximising a limited dataset, we use leave-one-out cross-validation during testing. This ensures that no forecast benefits from knowing the outcome it is trying to predict.

Each test mimics the real-world challenge of forecasting the future based only on the past. The separation between training and testing ensures the models are accurate and effective.

**Example:** We have harvest data from 2000 to 2025. To test how well the model predicts the yield in 2008, we train it on data from 2000 to 2025 excluding 2008. Then we predict 2008. This process is repeated for each year.

## Real Data

Our process replicates real forecasting conditions, where each prediction is made using only the information that would have been available *at the time*, so we can evaluate how well the model performs on genuinely unseen data—just as it would in live deployment.

**Example:** If we are testing if a model ran in June 2008, predicts the harvest in December 2008 (6 months early) we only used the data available in up to June 2008

## Point in Time

We maintain point-in-time integrity—preserving exactly what the model predicted at each point in history, using only the information available at that point in time.

Our predictions are then stored with a timestamp and never changed—even if new data shows the prediction was wrong.

All of our data is stored either daily or weekly. This gives us a full week-by-week record of past predictions, frozen in time.

This means that we can ask "*What did the model predicting in week 25 of 2008? How did that change in week 26?*". Crucially we can also ask "*How close were those forecasts to the final outcome?*"

The history ensures effective back testing and signal evaluation.

## Appendix 1: Data. Delivery. File Formats.

We provide three primary forecast metrics

- Area: Estimated planted or harvested hectares
- Yield: Crop output per hectare
- Production: Total output

| Data Type | Update schedule | Channels | Scope | History |
|---|---|---|---|---|
| Area | Weekly Sunday 19:00 to 23:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present / Predictive | 20+ Years |
| Yield | Weekly Sunday 19:00 to 23:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present / Predictive | 20+ Years |
| Production | Weekly Sunday 19:00 to 23:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present / Predictive | 20+ Years |
| Weather | Daily 05:00 to 07:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present / Predictive | 20+ Years |
| NDVI | Monthly Last day of each month 19:00 to 23:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present | 20+ Years |
| Crop Intelligence Reports | Weekdays 19:00 to 23:00 | AWS S3 bucket push, REST API pull, emailed files, interactive dashboards, scheduled PDF reports. | Past / Present | 1+ Year |

## Appendix 2: Crop Coverage

We provide data for a wide range for crops including, those listed below, Coverage is tailored by crop and region. Any crop or region can be added based on client requirements.

### Grains and Cereals

Comprehensive modelling is available for key global staples such as wheat, maize (corn), barley, and rice.

### Oilseeds

Extensive datasets are maintained for oilseed crops including soybeans, canola (rapeseed), sunflower, and palm oil.

### Soft Commodities

Sugar, coffee, cotton and cocoa, with detailed data models by country, region and even sub region (for example specific models for west Texas).

### Nuts

Support is available for a growing number of nut crops, including cashews, almonds, pistachios, macadamias, and groundnuts (peanuts). These models often rely on climate sensitivity indicators, orchard health assessments, and regional cost structures.

### Fruit Crops

Forecasting and monitoring are also provided for selected fruit crops, such as blueberries, raspberries, strawberries, peaches, passion fruit, grapes, and bananas. Data availability varies by crop and geography, with options for time-series weather data, satellite-derived phenology, and market-linked reporting.

### Root and Tuber Crops

Selected models are provided for crops like potatoes, sweet potatoes, and butternut squash,

### Legumes and Pulses

Support for beans, peas, sugar snap, and baby corn is available in selected regions,

### Horticultural and Specialty Crops

HSAT supports a range of specialty crops including asparagus, aubergines, courgettes, chilli peppers, pumpkins, sweetcorn, and tomatoes.