Remote Sensing Data Pipeline Documentation

Faiz Ahmad, fxa230012@utdallas.edu

January 30, 2026

A Agricultural Remote Sensing Data Schema

This document details the schema for the multi-sensor agricultural dataset generated via the Google Earth Engine (GEE) pipeline. The dataset aggregates satellite imagery, land cover classifications, and environmental variables into county-level annual statistics.

Table 1: Dataset Technical Specifications

Parameter	Specification
Spatial Granularity	U.S. Counties (FIPS)
Temporal Coverage	2000–2024 (Sensor dependent)
Aggregation Method	`reduceRegions` (Area-weighted)
Statistical Aggregators	Mean, StdDev, Percentiles (p10, p25, p50, p75, p90)
Categorical Handling	Normalized Class Percentages (Sum = 100% per county-year)
Missing Data	`NaN` indicates no valid pixels (clouds/sensor failure) or class absence

A.1 Common Identifier Columns

These columns are present in every CSV file and serve as the primary keys for joining datasets.

Table 2: Geospatial Identifiers

Dataset Variable Name	Standard Label	Definition	Units
`GEOID`	FIPS Code	Unique 5-digit county identifier (State + County)	–
`STATEFP`	State FIPS	2-digit state code	–
`COUNTYFP`	County FIPS	3-digit county code	–
`NAME`	County Name	Legal name of the county	–
`year`	Year	Calendar year of observation	Year

A.2 Dataset Composition Summary

Table 3 provides a comprehensive inventory of the multi-modal environmental data acquired for this study. The final dataset integrates ten distinct data streams, covering optical, radar, topographic, and agricultural domains, joined by unique county identifiers (FIPS/GEOID). There is significant overlap between Landsat 8/9 and Sentinel-2, which is excellent for dataset harmonization (often called the "virtual constellation" concept). The common features are - Optical Bands: Blue, Green, Red, NIR, SWIR 1, and SWIR 2. Indices: NDVI, EVI, NDMI, and BSI. While NDWI and SAVI can be calculated for Sentinel-2, they are not currently listed in our study. Despite the individual dataset's temporal and spatial resolution, we harmonized to 250 m using GEE - please see Appendix B for more on harmonization.

Table 3: Master Dataset Inventory & Source Specifications

Dataset Domain	Primary Source (Availability)	Key Variables & Feature Depth	Resolution & Frequency
Target Variable	IHME / CDC (2000–Present)	**1 Variable:** Mean Life Expectancy (Years)	Annual
Agriculture	USDA CDL (2008–Present)	**130+ Columns:** % Cover for major crops (Corn, Soy, Cotton, Rice) plus 100+ specific land cover classes.	Annual (30m)
Optical Imagery	Landsat 8/9 (2013–Present)	**84 Columns:** Surface Reflectance (Bands 2-7) + Indices (NDVI, EVI, NDWI, SAVI, BSI, NDMI) with full stats (mean, std, p10-p90).	Annual Comp. (30m)
Multi-Spectral	Sentinel-2 (2015–Present)	**84 Columns:** High-res bands (B2, B3, B4, B8, B11, B12) + Indices. Captures fine-scale vegetation health.	Annual Comp. (10m)
Radar Structure	Sentinel-1 SAR (2014–Present)	**42 Columns:** Backscatter Intensity (VV, VH) + GLCM Texture Metrics (Contrast, Entropy, Corr, ASM).	Annual Median (10m)
Temperature	MODIS (Terra) (2000–Present)	**14 Columns:** Land Surface Temperature (LST) for Day and Night (mean, std, percentiles).	8-Day Comp. (1km)
Vegetation Trend	MODIS Vegetation (2000–Present)	**14 Columns:** Regional-scale NDVI and EVI statistics for broad-scale greenness trends.	16-Day Comp. (250m)
Surface Water	JRC Global Water (1984–Present)	**4 Columns:** % Cover of Permanent Water, Seasonal Water, Non-Water, and No-Data pixels.	Annual (30m)
Topography	Copernicus DEM (Static ∼2015)	**7 Columns:** Elevation statistics (Mean, StdDev, p10-p90) defining terrain roughness.	Static (30m)
Soil Properties	OpenLandMap (Static)	**7 Columns:** Statistical aggregation (Mean, StdDev, p10-p90) of soil texture indices.	Static (250m)

B Individual Dataset Column Catalog

B.1 Landsat 8/9 Surface Reflectance

Source: NASA/USGS (`LC08/LC09`)
Resolution: 30m
Processing: Cloud masked, scaled to [0,1] surface reflectance, annual median composite.

Table 4: Landsat Spectral Bands and Indices

Dataset Variable Pattern	Standard Label	Definition	Units
Spectral Bands			
`SR_B2_{stat}` †	Blue Band	Surface reflectance $\lambda \approx 450 - 510$ nm	Unitless [0-1]
`SR_B3_{stat}` †	Green Band	Surface reflectance $\lambda \approx 530 - 590$ nm	Unitless [0-1]
`SR_B4_{stat}` †	Red Band	Surface reflectance $\lambda \approx 640 - 670$ nm	Unitless [0-1]
`SR_B5_{stat}` †	Near-Infrared (NIR)	Surface reflectance $\lambda \approx 850 - 880$ nm	Unitless [0-1]
`SR_B6_{stat}` †	SWIR 1	Shortwave Infrared $\lambda \approx 1570 - 1650$ nm	Unitless [0-1]
`SR_B7_{stat}` †	SWIR 2	Shortwave Infrared $\lambda \approx 2110 - 2290$ nm	Unitless [0-1]
Derived Indices			
`NDVI_{stat}` †	NDVI	Normalized Difference Vegetation Index: $\frac{NIR-Red}{NIR+Red}$	Unitless [-1,1]
`EVI_{stat}` †	EVI	Enhanced Vegetation Index (Atmosphere corrected)	Unitless [-1,1]
`SAVI_{stat}`	SAVI	Soil-Adjusted Vegetation Index ($L = 0.5$)	Unitless [-1,1]
`NDWI_{stat}`	NDWI	Normalized Difference Water Index: $\frac{Green-NIR}{Green+NIR}$	Unitless [-1,1]
`NDMI_{stat}` †	Moisture Index	Normalized Difference Moisture Index: $\frac{NIR-SWIR1}{NIR+SWIR1}$	Unitless [-1,1]
`BSI_{stat}` †	Bare Soil Index	Index enhancing bare soil identification	Unitless [-1,1]

Note: {stat} represents aggregation statistics (mean, stdDev, p10-p90).
† *Indicates features harmonizable with Sentinel-2.*

B.2 Sentinel-2 Multi-Spectral Instrument

Source: ESA Copernicus (`S2_SR_HARMONIZED`)
Resolution: 10m/20m
Processing: Cloud masked via QA60, scaled to [0,1], annual median composite.

Table 5: Sentinel-2 Spectral Features

Dataset Variable Pattern	Standard Label	Definition	Units
`B2_{stat}` †	Blue Reflectance	Band 2 (490 nm)	Unitless
`B3_{stat}` †	Green Reflectance	Band 3 (560 nm)	Unitless
`B4_{stat}` †	Red Reflectance	Band 4 (665 nm)	Unitless
`B8_{stat}` †	NIR Reflectance	Band 8 (842 nm)	Unitless
`B11_{stat}` †	SWIR 1 Reflectance	Band 11 (1610 nm)	Unitless
`B12_{stat}` †	SWIR 2 Reflectance	Band 12 (2190 nm)	Unitless
`NDVI_{stat}`, `EVI_{stat}` †	Vegetation Indices	See Landsat definitions	Unitless
`NDMI_{stat}`, `BSI_{stat}` †	Environmental Indices	See Landsat definitions	Unitless

† *Indicates features harmonizable with Landsat 8/9.*

B.3 Sentinel-1 SAR (Radar)

Source: ESA Copernicus (`S1_GRD`)
Mode: Interferometric Wide (IW), Ground Range Detected (GRD)
Processing: Thermal noise removal, radiometric calibration, terrain correction.

Table 6: Sentinel-1 Backscatter and Texture Metrics

Dataset Variable Pattern	Standard Label	Definition	Units
Backscatter Intensity			
`VV_{stat}`	Vertical Backscatter	Vertical Transmit/Receive intensity	dB
`VH_{stat}`	Cross-Pol Backscatter	Vertical Transmit/Horizontal Receive intensity	dB
GLCM Texture Features (Calculated on VV Band)			
`VV_asm_{stat}`	Texture Uniformity	Angular Second Moment (measure of homogeneity)	Unitless
`VV_contrast_{stat}`	Texture Contrast	Measure of local intensity variation	Unitless
`VV_corr_{stat}`	Texture Correlation	Linear dependency of gray levels	Unitless
`VV_ent_{stat}`	Texture Entropy	Randomness/disorder (proxy for vegetation complexity)	Unitless

B.4 MODIS Products

Source: NASA Terra Satellite (`MOD13A1`, `MOD11A2`)
Temporal Depth: 2000–Present

Table 7: MODIS Vegetation and Temperature

Dataset Variable Pattern	Standard Label	Definition	Units
`NDVI_{stat}`	MODIS NDVI	Regional vegetation greenness	Unitless
`EVI_{stat}`	MODIS EVI	Regional enhanced vegetation index	Unitless
`LST_Day_1km_{stat}`	Daytime LST	Mean daytime Land Surface Temperature	Celsius (°C)
`LST_Night_1km_{stat}`	Nighttime LST	Mean nighttime Land Surface Temperature	Celsius (°C)

B.5 Categorical Land Cover Features

All categorical variables are normalized as percentages of the county area: $\sum(\text{Classes}) = 100\%$.

B.6 USDA Cropland Data Layer (CDL)

Table 8: USDA Agricultural Commodities

Dataset Variable Name	Standard Label	Definition	Units
Cropland_USDA_Corn_pct	Corn	Class 1: Corn	%
Cropland_USDA_Cotton_pct	Cotton	Class 2: Cotton	%
Cropland_USDA_Rice_pct	Rice	Class 3: Rice	%
Cropland_USDA_Sorghum_pct	Sorghum	Class 4: Sorghum	%
Cropland_USDA_Soybeans_pct	Soybeans	Class 5: Soybeans	%
Cropland_USDA_Sunflower_pct	Sunflower	Class 6: Sunflower	%
Cropland_USDA_Peanuts_pct	Peanuts	Class 10: Peanuts	%
Cropland_USDA_Tobacco_pct	Tobacco	Class 11: Tobacco	%
Cropland_USDA_Sweet Corn_pct	Sweet Corn	Class 12: Sweet Corn	%
Cropland_USDA_Barley_pct	Barley	Class 21: Barley	%
Cropland_USDA_Spring Wheat_pct	Spring Wheat	Class 23: Spring Wheat	%
Cropland_USDA_Winter Wheat_pct	Winter Wheat	Class 24: Winter Wheat	%
Cropland_USDA_Oats_pct	Oats	Class 28: Oats	%
Cropland_USDA_Alfalfa_pct	Alfalfa	Class 36: Alfalfa	%
Cropland_USDA_Other Hay_pct	Other Hay	Class 37: Non-Alfalfa Hay	%
Cropland_USDA_Fallow_pct	Fallow Land	Class 61: Idle Cropland	%

Note: The dataset also contains raw class columns (e.g., Cropland_USDA_Class_88_pct) for non-major crops and non-agricultural land covers, which are omitted here for brevity.

B.7 JRC Global Surface Water

Table 9: JRC Water Dynamics

Dataset Variable Name	Standard Label	Definition	Units
Water_JRC_NotWater_pct	Land Coverage	Areas never detected as water	%
Water_JRC_Seasonal_pct	Seasonal Water	Areas with water $<$ 12 mo/yr	%
Water_JRC_Permanent_pct	Permanent Water	Areas with water 12 mo/yr	%
Water_JRC_Class_0_pct	No Data / Masked	Invalid pixels or no observation	%

B.8 Topography and Soil

Table 10: Static Environmental Features (DEM & Soil)

Dataset Variable Pattern	Standard Label	Definition	Units
Copernicus DEM (GLO-30)			
`DEM\_mean`	Mean Elevation	Average elevation above sea level	Meters
`DEM\_stdDev`	Terrain Roughness	Standard deviation of elevation	Meters
`DEM\_pXX`	Elevation Percentile	Distribution ($p10$–$p90$)	Meters

Note: Soil variables represent statistical aggregations of the underlying texture class indices.

C Data Acquisition & Validation Strategy

The following sections document the Earth Engine (GEE) sourcing and spatial reduction strategies used to generate the master dataset. All logic is verified against the project's source code configuration.

C.1 Aggregated Validation Summary

Data accumulation was performed using a Python-GEE pipeline with two distinct reduction strategies determined by the asset configuration:

- **Continuous Reducers:** Applied to Landsat, Sentinel, MODIS, DEM, and Soil.

  `ee.Reducer.mean().combine(ee.Reducer.stdDev()).combine(ee.Reducer.percentile([10,25,50`

- **Categorical Reducers:** Applied to USDA Cropland and JRC Water.

  `ee.Reducer.frequencyHistogram()` (Post-processed to percentage columns).

C.2 Dataset-Specific Sourcing Verification

C.2.1 Landsat 8/9 Surface Reflectance

- **GEE Asset:** `LANDSAT/LC08/C02/T1_L2` (merged with LC09).
- **Processing Code:**

```
# Cloud masking and Scaling
image.updateMask(mask).multiply(0.0000275).add(-0.2)
# Spectral Indices added via add_spectral_indices():
# NDVI, EVI, SAVI, NDWI, NDMI, BSI
```

- **Validation of Columns:** The pipeline applies continuous reducers to all optical bands and computed indices.
- **Output Columns:** `SR_B[2-7]_mean`, `NDVI_p50`, `NDMI_stdDev`, etc.

C.2.2 Sentinel-2 MSI

- **GEE Asset:** `COPERNICUS/S2_SR_HARMONIZED`.
- **Processing Code:**

```
# QA60 bitmasking for clouds/cirrus
image.updateMask(mask).divide(10000)
# Indices: Same function as Landsat (NDVI, EVI, etc.)
```

- **Validation of Columns:** Continuous reduction confirms presence of statistical distributions for spectral bands.
- **Output Columns:** `B[2,3,4,8,11,12]_mean`, `EVI_p90`, etc.

C.2.3 Sentinel-1 SAR

- **GEE Asset:** `COPERNICUS/S1_GRD`.
- **Processing Code:**

```
# Texture feature generation (GLCM)
glcm = img.select(['VV']).glcmTexture(size=3)
texture_bands = ['VV_asm', 'VV_contrast', 'VV_corr', 'VV_ent']
```

- **Validation of Columns:** Code explicitly adds GLCM bands to the reducer list.
- **Output Columns:** `VV_mean`, `VH_p50`, `VV_contrast_mean`, etc.

C.2.4 MODIS Products (Vegetation & Temperature)

- **GEE Assets:**
  - NDVI: `MODIS/061/MOD13A1`
  - LST: `MODIS/061/MOD11A2`
- **Processing Code (LST):**

```
img.multiply(0.02).subtract(273.15) # Kelvin to Celsius conversion
```

- **Validation of Columns:** Configuration flags `categorical` were set to `False`, triggering continuous statistics.
- **Output Columns:** `LST_Day_1km_mean`, `NDVI_p50`, `EVI_stdDev`.

C.2.5 USDA Cropland Data Layer (CDL)

- **GEE Asset:** `USDA/NASS/CDL`.
- **Processing Strategy:**

```
'categorical': True
# Triggers: red.combine(ee.Reducer.frequencyHistogram())
```

- **Validation of Columns:** The histogram output (JSON) was "exploded" during post-processing to create columns for each crop class code.
- **Output Columns:** `Cropland_USDA_Corn_pct`, `Class_88_pct`, etc.

C.2.6 JRC Global Surface Water

- **GEE Asset:** `JRC/GSW1_4/YearlyHistory`.
- **Processing Strategy:** `'categorical': True`.
- **Validation of Columns:** Histogram reduction preserves pixel counts for water classes.
- **Output Columns:** `Water_JRC_Permanent_pct`, `Water_JRC_Seasonal_pct`.

C.2.7 Digital Elevation Model

- **GEE Asset:** `COPERNICUS/DEM/GLO30`.
- **Processing Strategy:** Static asset reduced via continuous statistics.
- **Validation of Columns:** Code confirms generation of distributional stats for elevation.
- **Output Columns:** `DEM_mean`, `DEM_p90`, `DEM_stdDev`.

C.2.8 Soil Properties (OpenLandMap)

- **GEE Asset:** `OpenLandMap/SOL/SOL_TEXTURE-CLASS_USDA-TT_M/v02`.
- **Important Validation Note:** Although the source data represents categorical texture classes (Integers 1-12), the configuration dictionary in the pipeline defined:

```
'soil_texture': { ..., 'static': True }
# Missing 'categorical': True
```

  Consequently, the pipeline applied the **continuous reducer** path (Mean, Percentiles) rather than the histogram path. This confirms why the output CSV contains statistical metrics rather than class percentages.

- **Output Columns:** `mean`, `p10`, `p90`, `stdDev` (Generated on texture class IDs).

- - Thank you - -