

# CEWP Data Source Audit

## Spatio-Temporal Resolutions & Imputation Methods

Pipeline Version	December 2025
Target Region	Central African Republic (CAR)
Analysis Grid	H3 Resolution 5 (~10km hexagonal cells)
Grid Coverage	3,407 cells
Temporal Spine	14-day intervals aligned to 2000-01-01
Generated	2025-12-08 19:07

# Executive Summary

This audit documents all data sources integrated into the Conflict Early Warning Pipeline (CEWP), detailing their native resolutions, pipeline transformations, and imputation strategies for missing values.

Category	Sources	Native Resolution	Target Resolution
Environmental	7	100m – 5km, hourly/daily	H3-5, 14-day
Conflict & Events	3	Point, daily	H3-5, 14-day
Socio-Political	3	Admin-1, annual	H3-5, 14-day
Economic	2	National, daily	National, 14-day
Infrastructure	5	Point/polygon, static	H3-5, static
Demographics	1	100m, annual	H3-5, annual

Table 1: Data source categories with resolution transformations

## 1. Environmental Data

Environmental variables are aggregated from Google Earth Engine collections to H3 resolution 5 with 14-day temporal windows. Server-side processing reduces data transfer and enables efficient historical analysis.

### 1.1 CHIRPS Precipitation

Source	UCSB-CHG/CHIRPS/DAILY (via GEE)
Native Spatial	0.05° (~5.5km)
Native Temporal	Daily
Aggregation	H3-5 zonal mean, 14-day sum
Output Column	precip_mean_depth_mm
Imputation	Forward-fill (limit=4 steps / 56 days)
Coverage	1981–present

### 1.2 ERA5-Land Climate

Source	ECMWF/ERA5_LAND/HOURLY (via GEE)
Native Spatial	0.1° (~11km)
Native Temporal	Hourly
Variables	Temperature (2m), Dewpoint (2m), Soil Moisture (Layer 1)

<b>Aggregation</b>	H3-5 zonal mean, 14-day mean
<b>Output Columns</b>	temp_mean, dew_mean, soil_moisture_mean
<b>Imputation</b>	Forward-fill (limit=4 steps)
<b>Coverage</b>	1950–present (~5-day lag)

### 1.3 MODIS Vegetation (NDVI)

<b>Source</b>	MODIS/061/MCD43A4 (via GEE)
<b>Native Spatial</b>	500m
<b>Native Temporal</b>	Daily (16-day composite)
<b>Aggregation</b>	H3-5 zonal max, 14-day max
<b>Output Column</b>	ndvi_max
<b>Imputation</b>	Forward-fill (limit=4 steps)
<b>Notes</b>	NDVI = (NIR - Red) / (NIR + Red); range [-1, 1]

### 1.4 VIIRS Nighttime Lights

<b>Source</b>	NASA/VIIRS/002/VNP46A2 (via GEE)
<b>Native Spatial</b>	500m
<b>Native Temporal</b>	Daily
<b>Aggregation</b>	H3-5 zonal mean, 14-day mean
<b>Output Column</b>	ntl_mean
<b>Imputation</b>	Forward-fill (limit=4 steps)
<b>Coverage</b>	2012–present
<b>Notes</b>	DNB_BRDF_Corrected_NTL band; pre-2012 years have no data

### 1.5 JRC Global Surface Water

<b>Source</b>	JRC/GSW1_4/MonthlyHistory (pre-2022) or Landsat 8/9 (2022+)
<b>Native Spatial</b>	30m
<b>Native Temporal</b>	Monthly (JRC) / Daily composite (Landsat)
<b>Aggregation</b>	H3-5 zonal mean + max, 14-day
<b>Output Columns</b>	water_local_mean, water_local_max
<b>Imputation</b>	Forward-fill (limit=4 steps)

**Notes**

Binary water detection (MNDWI > 0.1 for Landsat)

## 2. Conflict & Event Data

Event data is aggregated from point locations to H3 cells with 14-day temporal windows. Missing values are zero-filled under the assumption that no recorded event equals no event occurrence.

### 2.1 ACLED (Armed Conflict Location & Event Data)

<b>Source</b>	Local CSV (data/raw/acled.csv)
<b>Native Spatial</b>	Point (lat/lon, geo_precision 1-3)
<b>Native Temporal</b>	Daily (event_date)
<b>Aggregation</b>	H3-5 point-in-polygon, 14-day sum
<b>Output Columns</b>	fatalities_14d_sum, protest_count, riot_count, violence_event_count
<b>Imputation</b>	Zero-fill for missing (no events = 0)
<b>Coverage</b>	1997–present (CAR: 2000+)
<b>Volume</b>	7,584 events in CAR

### 2.2 GDELT (Global Database of Events, Language, and Tone)

<b>Source</b>	BigQuery gdelt-bq.gdeltv2.events
<b>Native Spatial</b>	Point (Actor1Geo_Lat/Lon)
<b>Native Temporal</b>	Daily (SQLDATE)
<b>Aggregation</b>	H3-5 point-in-polygon, 14-day sum/mean
<b>Output Columns</b>	gdelt_event_count, gdelt_goldstein_mean, gdelt_mentions_total, gdelt_avg_tone
<b>Imputation</b>	Zero-fill for missing
<b>Coverage</b>	2015–present (v2)
<b>Volume</b>	162,116 feature rows for CAR
<b>Notes</b>	Uses FIPS code "CT" for CAR

### 2.3 IODA (Internet Outage Detection & Analysis)

<b>Source</b>	Georgia Tech IODA API
<b>Native Spatial</b>	Country or Admin-1 region
<b>Native Temporal</b>	Event-based (start/end timestamps)
<b>Aggregation</b>	Admin-1 → H3-5 polygon overlay, daily binary
<b>Output Column</b>	ioda_outage_detected

<b>Imputation</b>	Zero-fill (no outage = 0)
<b>Coverage</b>	2016–present
<b>Volume</b>	798,015 H3-daily records
<b>Notes</b>	Disaggregated via Admin-1 shapefile

### 3. Socio-Political Data

Socio-political data requires spatial disaggregation from administrative boundaries to H3 cells. These sources capture ethnic power dynamics, displacement patterns, and food security conditions.

#### 3.1 EPR (Ethnic Power Relations)

Source	ETH Zurich EPR-2021 (CSV + GeoJSON)
Native Spatial	Ethnic group polygon (GeoEPR)
Native Temporal	Annual (from-to year ranges)
Aggregation	Polygon → H3-5 via <code>polygon_to_cells()</code> , annual
Output Columns	<code>ethnic_group_count</code> , <code>epr_excluded_groups_count</code> , <code>epr_status_mean</code> , <code>epr_horizontal_inequality</code>
Imputation	Zero-fill for cells outside any ethnic polygon
Coverage	1946–2021 (expanded year-by-year)
Volume	27 polygons in CAR; 282 group-year records

#### 3.2 IOM DTM (Displacement Tracking Matrix)

Source	IOM DTM API v3
Native Spatial	Admin-2 (sub-prefecture) or Admin-1
Native Temporal	Survey rounds (irregular, ~quarterly)
Aggregation	Admin-2 → H3-5 via area-weighted density, 14-day
Output Column	<code>iom_displacement_sum</code>
Imputation	Zero-fill between survey rounds
Coverage	2014–present
Volume	3,735 Admin-2 records → 40,141 H3 records
Notes	Fallback to Admin-1 if Admin-2 unavailable

#### 3.3 FEWS NET IPC (Food Security Phases)

Source	FEWS NET Data Warehouse API
Native Spatial	Admin-1 (prefecture)
Native Temporal	Quarterly projections
Aggregation	Admin-1 → H3-5 via polygon overlay, 14-day max
Output Column	<code>ipc_phase_class</code> (1-5 scale)

<b>Imputation</b>	Constant=0 before 2009-01-01 (pre-IPC era); forward-fill after
<b>Coverage</b>	2009–present
<b>Notes</b>	Requires FEWS_NET_TOKEN in .env

## 4. Economic Data

Economic indicators are non-spatial (global commodities) or market-specific (food prices). These are joined to all H3 cells uniformly or via market proximity.

### 4.1 Yahoo Finance Commodities

<b>Source</b>	yfinance Python package
<b>Tickers</b>	GC=F (Gold), CL=F (Oil), ^GSPC (S&P 500), EURUSD=X
<b>Native Spatial</b>	Global (non-spatial)
<b>Native Temporal</b>	Daily (trading days)
<b>Aggregation</b>	National-level, 14-day mean
<b>Output Column</b>	commodity_gold_price_usd
<b>Imputation</b>	Forward-fill for non-trading days
<b>Coverage</b>	2000-08-30 (Gold futures) to present
<b>Volume</b>	69,987 records uploaded

### 4.2 FEWS NET Market Prices

<b>Source</b>	FEWS NET Data Warehouse API
<b>Native Spatial</b>	Market locations (10 markets in CAR)
<b>Native Temporal</b>	Monthly
<b>Aggregation</b>	Market → nearest H3-5 cell, 14-day
<b>Output Columns</b>	Market-specific price features
<b>Imputation</b>	Forward-fill
<b>Coverage</b>	2015–present
<b>Notes</b>	Requires FEWS_NET_TOKEN



## 5. Infrastructure & Geography (Static)

Static infrastructure features are computed once and joined to all temporal observations. Distance calculations use cKDTree for efficient nearest-neighbor queries.

Data Source	Native Resolution	Output Features	Notes
Copernicus DEM	90m raster	elevation_mean, slope_mean, terrain_ruggedness_mean	via Sentinel Hub
GRIP4 Roads	Polyline network	dist_to_road (km)	PBL Region 3 (Africa)
HydroRIVERS	Polyline network	dist_to_river (km)	Stream order $\geq 3$ ; 8,811 segments
IPIS Mining Sites	Point locations	dist_to_diamond_mine, dist_to_gold_mine (km)	914 sites $\rightarrow$ 125 H3 cells
OSM Settlements	Point locations	dist_to_city, dist_to_capital (km)	Via HDX

Table 2: Static infrastructure data sources and derived features

## 6. Demographics

### 6.1 WorldPop Population

Source	WorldPop 100m constrained (R2025A)
Native Spatial	100m raster
Native Temporal	Annual
Aggregation	H3-5 zonal sum, annual
Output Columns	pop_count, pop_log
Imputation	Forward-fill within year; backward extrapolation (2.5% annual growth) pre-2015
Coverage	2000–2025
File Variants	caf_ppp_{year}_UNadj.tif (2000-2014), caf_pop_{year}_CN_100m_R2025A_v1.tif (2015+)

## 7. Imputation Strategy

Missing values are handled through feature-specific strategies that preserve temporal and spatial coherence while avoiding information leakage from future observations.

### 7.1 Default Strategy

The default imputation method is forward-fill with a 4-step limit (56 days maximum gap). This preserves the last known value while preventing stale data from persisting indefinitely.

```
imputation:
  defaults:
    method: "forward_fill"
    limit: 4 # 4 steps = 56 days max gap
```

### 7.2 Feature-Specific Overrides

Feature	Method	Details
Population	Backward extrapolation	2.5% annual growth rate, start_year=2015
IPC Phase	Constant	Value=0 before 2009-01-01 (pre-IPC era)
Conflict events	Zero-fill	No event = 0 fatalities/protests/riots
GDELT events	Zero-fill	No event = 0 count
IODA outages	Zero-fill	No outage = 0
IOM displacement	Zero-fill	Between survey rounds
EPR features	Zero-fill	Cells outside ethnic polygons

Table 3: Feature-specific imputation overrides

## 8. Resolution Transformations

All data sources are harmonized to a common spatio-temporal framework through aggregation or disaggregation. The following diagrams illustrate the transformation pathways.

### 8.1 Spatial Resolution Mapping

Native Resolution	Sources	Transformation
30m	Landsat water	Zonal statistics → H3-5
90m	Copernicus DEM	Zonal statistics → H3-5
100m	WorldPop	Zonal sum → H3-5

Native Resolution	Sources	Transformation
500m	MODIS, VIIRS	Zonal statistics → H3-5
5km	CHIRPS	Zonal mean → H3-5
11km	ERA5	Zonal mean → H3-5
Point	ACLED, GDELT, mines, settlements	Point-in-polygon → H3-5
Admin-1	IPC, IODA	Polygon overlay → H3-5
Admin-2	IOM	Area-weighted disaggregation → H3-5

*Table 4: Spatial resolution transformations to H3 Resolution 5 (~10km)*

## 8.2 Temporal Resolution Mapping

Native Frequency	Sources	Transformation
Hourly	ERA5	14-day mean
Daily	CHIRPS, ACLED, GDELT	14-day sum/mean
16-day composite	MODIS	14-day max
Monthly	JRC water, market prices	14-day interpolation
Quarterly	IOM, IPC	14-day forward-fill
Annual	WorldPop, EPR	Joined to all steps in year
Static	DEM, roads, rivers, mines	Joined to all time steps

*Table 5: Temporal resolution transformations to 14-day intervals*

## 9. Known Data Gaps & Issues

The following issues were identified during the pipeline audit and may affect model performance.

Issue	Impact	Status
FEWS_NET_TOKEN missing	IPC + market prices = NULL	■ User action required
Market locations CSV BOM	0 markets loaded	■ Code fix needed
VIIRS pre-2012	NTL = NULL	✓ Expected (sensor launch)
EPR post-2021	Uses 2021 data for 2022-2025	✓ Acceptable
IOM survey gaps	Irregular temporal coverage	✓ Zero-filled
Economy data query	Gold price = NULL	■ Code fix needed

Table 6: Known data gaps and their resolution status

## 10. Temporal Transformations

Raw features undergo temporal transformations to capture lagged effects, cumulative impacts, and anomalies relative to historical baselines.

Transformation	Description	Example Output
lag_1_step	Shift by 1 period (14 days)	temp_mean_lag1
sum_1_step	Sum within period	fatalities_14d_sum
decay_30d	Exponential decay (half-life ~30 days, span=2.14)	fatalities_decay_30d
anomaly_6_step	Value minus 6-period rolling mean	chirps_precip_anomaly
spatial_mean_k	Mean of k-ring neighbors	(defined but not active)

Table 7: Temporal transformation functions applied during feature engineering