CEWP Data Source Audit v4.0

Spatio-Temporal Resolutions & Imputation Methods

Pipeline Version	January 2026
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	2026-01-17

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. The pipeline harmonizes 21+ distinct data sources across seven thematic categories into a unified analytical framework suitable for conflict prediction modeling.

Category	Sources	Native Resolution	Target Resolution
Environmental	8	100m – 11km, hourly/daily	H3-5, 14-day
Conflict & Events	3	Point, daily	H3-5, 14-day
Socio-Political	4	Admin-1/2, annual/quarterly	H3-5, 14-day
Economic	6	National/Market, daily/monthly	National/H3-5, 14-day
Infrastructure	5	Point/polygon, static	H3-5, static
Demographics	1	100m, annual	H3-5, annual
NLP/Semantic	1	Event-level text	H3-5, 14-day

Temporal Lag Handling

The pipeline distinguishes between two independent lag mechanisms:

Publication Lags (Data Availability)

Publication lags account for the delay between data collection and public availability. These are applied at ingestion or immediately after merge so timestamps reflect when data would actually be available in a real-time deployment.

Source	Publication Lag	Applied At
GEE Environmental (ERA5, CHIRPS, MODIS, VIIRS, JRC, Dynamic World)	14 days (1 step)	Storage (GEE window shift)
ACLED Hybrid NLP Drivers	14 days (1 step)	Storage (event_date shift)
ACLED Conflict Counts	14 days (1 step)	Feature engineering (post-

		merge)
Food Prices (WFP/FEWS NET)	56 days (4 steps)	Storage (date shift)

Analytical Lags (Leakage Control)

Analytical lags ensure features only use prior-period values, preventing temporal leakage into the model. These are applied downstream in feature engineering via lag_1_step and decay transforms.

Mechanism	Purpose	Implementation
LAG() / shift(1)	Features use prior time step values	Applied to all model features
LEAD()	Targets reference future time steps	Applied to prediction targets
Rolling windows	Strictly historical aggregation	No overlap with target period

Key distinction: Publication lag moves the timestamp to when data is available; analytical lag shifts what the model sees to a prior window. A feature can have both—e.g., GEE data has a 14-day publication lag at ingestion AND an analytical lag when used as a model feature.

1. Environmental Data

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

Publication Lag: All GEE-sourced data uses window [T, T+13] but is stored/labeled at date T+14 to prevent temporal leakage.

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
Imputation	Forward-fill (limit=4 steps / 56 days)
Coverage	1981–present
Publication Lag	14 days (applied at storage)

Output Columns:

- precip_mean_depth_mm → chirps_precip_anomaly

1.2 ERA5-Land Climate

Source	ECMWF/ERA5_LAND/HOURLY (via GEE)
Native Spatial	0.1° (~11km)
Native Temporal	Hourly
Aggregation	H3-5 zonal mean, 14-day mean
Imputation	Forward-fill (limit=4 steps)
Coverage	1950–present (~5-day lag)
Publication Lag	14 days (applied at storage)

Variables:

- Temperature (2m)
- Dewpoint (2m)
- Soil Moisture (Layer 1)

Output Columns:

- era5_temp_anomaly
- era5_soil_moisture_anomaly

1.3 MODIS Vegetation (NDVI)

Source	MODIS/061/MCD43A4 (via GEE)
Native Spatial	500m

Native Temporal	Daily (16-day composite)
Aggregation	H3-5 zonal max, 14-day max
Imputation	Forward-fill (limit=4 steps)
Notes	NDVI = (NIR - Red) / (NIR + Red); range [-1, 1]
Publication Lag	14 days (applied at storage)

Output Columns:

- ndvi_max → ndvi_anomaly

1.4 VIIRS Nighttime Lights

Source	NASA/VIIRS/002/VNP46A2 (via GEE)
Native Spatial	500m
Native Temporal	Daily
Aggregation	H3-5 zonal mean, 14-day mean
Imputation	Forward-fill (limit=4 steps)
Coverage	2012–present
Structural Break	Pre-2012 data = NULL (sensor launch date)
Publication Lag	14 days (applied at storage)

Output Columns:

- ntl_mean → nightlights_intensity

1.5 JRC Global Surface Water

Source	JRC/GSW1_4/MonthlyHistory (pre-2022) or Landsat 8/9 (2022+)
Native Spatial	30m
Native Temporal	Monthly (JRC) / Daily composite (Landsat)
Aggregation	H3-5 zonal mean + max, 14-day
Imputation	Forward-fill (limit=4 steps)
Notes	Binary water detection (MNDWI > 0.1 for Landsat)
Publication Lag	14 days (applied at storage)

Output Columns:

- water_coverage_lag1
- water_presence_lag1

1.6 Dynamic World Landcover

Source	GOOGLE/DYNAMICWORLD/V1 (via GEE)
Native Spatial	10m
Native Temporal	Near real-time (Sentinel-2 derived)
Aggregation	H3-5 zonal mean, 14-day mean probability
Availability Flag	landcover_avail: 1 if date ≥ 2015-06-27, else 0
Imputation	Forward-fill (limit=4 steps)
Coverage	2015-06-27–present
Publication Lag	14 days (applied at storage)

Output Columns (all as mean fraction 0-1):

- landcover_grass
- landcover_crops
- landcover_trees
- landcover_bare
- landcover_built

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)

Source	Local CSV (data/raw/ACLED.csv)
Native Spatial	Point (lat/lon, geo_precision 1-3)
Native Temporal	Daily (event_date)
Aggregation	H3-5 point-in-polygon, 14-day sum
Imputation	Zero-fill for missing (no events = 0)
Coverage	1997–present (CAR: 2000+)
Volume	~7,500+ events in CAR
Publication Lag	14 days (applied to features, not targets)

Output Columns:

- fatalities_14d_sum
- fatalities_1m_lag
- conflict_density_10km
- protest_count_lag1
- riot_count_lag1
- regional_risk_score_lag1

Note: ACLED features and targets are handled separately. Features receive 14-day publication lag; targets remain unlagged to preserve prediction validity.

2.2 ACLED Hybrid NLP Features

Semi-supervised semantic projection approach using ensemble scoring and residual clustering.

Source	ACLED notes field processed by process_acled_hybrid.py
Native Spatial	Event-level (inherits H3 from parent event)
Native Temporal	Event date
Embedding Model	all-MiniLM-L6-v2 (384-dimensional)
Method	Hybrid: 8 semantic themes (residual clustering) + 5 regex/anchor drivers
Scoring	Ensemble: regex pattern matching + cosine similarity to anchor phrases
Output Type	Continuous risk scores (0.0-1.0)
Imputation	Zero-fill for missing
Coverage	Full ACLED history

Total Features	13 (8 themes + 5 drivers)
Publication Lag	14 days (shifted at storage)

Theme Columns (8):

- theme_context_0
- theme_context_1
- theme_context_2
- theme_context_3
- theme_context_4
- theme_context_5
- theme_context_6
- theme_context_7

2.2.1 Explicit Drivers (5 total)

Driver	Description
driver_resource_cattle	Cattle/livestock-related conflict (raiding, grazing disputes)
driver_resource_mining	Mining-related violence (artisanal mining, resource extraction)
driver_econ_taxation	Economic extortion (illegal taxation, checkpoint fees)
driver_political_coup	Political instability indicators (coup attempts, power struggles)
driver_civilian_abuse	Human rights violations (civilian targeting, abuse)

2.3 GDELT (Global Database of Events, Language, and Tone)

Source	BigQuery gdelt-bq.gdeltv2.events
Native Spatial	Point (Actor1Geo_Lat/Lon)
Native Temporal	Daily (SQLDATE)
Aggregation	H3-5 point-in-polygon, 14-day sum/mean
Imputation	Zero-fill for missing
Coverage	2015–present (v2)
Notes	Uses FIPS code "CT" for CAR

Output Columns:

- events_3m_lag
- gdelt_decay_30d
- gdelt_avg_tone_decay_30d

3. Socio-Political Data

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 polygon_to_cells(), annual
Imputation	Zero-fill for cells outside any ethnic polygon
Coverage	1946–2021 (expanded year-by-year)
Post-2021	Uses 2021 data (acceptable for slow-changing features)

Output Columns:

- ethnic_group_count
- epr_excluded_groups_count
- epr_discriminated_groups_count
- epr_status_mean

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
Imputation	Zero-fill between survey rounds
Coverage	2015-01-31–present
Notes	Fallback to Admin-1 if Admin-2 unavailable

Output Columns:

- iom_displacement_count_lag1
- iom_data_available (structural break flag)

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
Imputation	Constant=0 before 2009-01-01 (pre-IPC era); forward-fill after

Coverage	2009–present

Output Columns:

- ipc_phase_class (1-5 scale)

3.4 IODA Internet Outage Detection

Source	Georgia Tech IODA API (outage events endpoint only)
Native Spatial	National (country-level; regional data not available for CAR)
Native Temporal	Event-based (near real-time outage detection)
Aggregation	National → broadcast to all H3 cells, 14-day weighted sum
Imputation	Zero-fill (no outage = 0)
Coverage	2022-02-01–present
Notes	Outage events only; no connectivity index feed. Higher score = more/worse outages.

Output Columns:

- ioda_outage_score
- ioda_data_available

4. Economic Data

4.1 Macro-Economic Indicators

Source	Yahoo Finance (yfinance package)
Tickers	GC=F (Gold), CL=F (Oil), ^GSPC (S&P 500), EURUSD=X
Native Spatial	Global (non-spatial, broadcast to all cells)
Native Temporal	Daily (trading days)
Aggregation	National-level, 14-day mean
Imputation	Forward-fill for non-trading days
Coverage	2003-12-01–present (code clamps start date)
Structural Break	econ_data_available = 0 before 2003-12-01, else 1

Output Columns:

- gold_price_usd_lag1
- oil_price_usd_lag1
- sp500_index_lag1
- eur_usd_rate_lag1

4.2 WFP/FEWS NET Food Prices

Source	FEWS NET Data Warehouse API
Native Spatial	Market locations (10+ markets in CAR)
Native Temporal	Monthly
Aggregation	Market → nearest H3-5 cell, 14-day
Imputation	Forward-fill
Coverage	2015–present
Publication Lag	56 days (4 steps) shifted at storage
Notes	Food price index averages only available, non-zero commodities

Output Columns:

- price_maize
- price_rice
- price_oil
- price_sorghum
- food_price_index

Shock Features:

- price_maize_shock
- price_rice_shock
- price_oil_shock
- price_sorghum_shock

5. Infrastructure & Geography (Static)

Data Source	Native Resolution	Output Features
Copernicus DEM	90m raster	elevation_mean slope_mean terrain_ruggedness_index
GRIP4 Roads	Polyline network	dist_to_road (km)
HydroRIVERS	Polyline network	dist_to_river (km)
IPIS Mining Sites	Point locations	dist_to_diamond_mine dist_to_gold_mine dist_to_large_mine dist_to_controlled_mine dist_to_large_gold_mine
OSM Settlements	Point locations	dist_to_city dist_to_capital (km)
CAR Boundary	Polygon	dist_to_border (km)

6. Demographics

6.1 WorldPop Population

Source	WorldPop 100m constrained (R2025A)
Native Spatial	100m raster
Native Temporal	Annual
Aggregation	H3-5 zonal sum, annual
Imputation	Forward-fill within each hex; pre-coverage gaps set to 0 (no backward extrapolation)
Transform	log1p(pop_count) → pop_log
Coverage	2000–2030
Structural Break	2015: V1 (census-adjusted) → V2 (constrained model), flagged via is_worldpop_v1

Output Columns:

- pop_count
- pop_log
- is_worldpop_v1

7. Temporal Context Features

Feature	Description
month_sin	Sine of month (cyclical seasonality)
month_cos	Cosine of month (cyclical seasonality)

is_dry_season	Binary: 1 if Nov-Mar

8. Structural Break Handling

Break Flag	Threshold Date	Purpose
is_worldpop_v1	Pre-2015	Census-adjusted vs constrained population
iom_data_available	Pre-2015-01-31	IOM DTM data start
econ_data_available	Pre-2003-12-01	Yahoo Finance coverage start
ioda_data_available	Pre-2022-02-01	IODA monitoring start
landcover_avail	Pre-2015-06-27	Dynamic World start

9. Feature Count Summary

Total Features: 111 (45 raw + 66 transformed)

Category	Raw	Transformed	Total
Environmental	10	16	26
Conflict	5	15	20
ACLED Hybrid NLP	0	13	13
Economic	8	12	20
Socio-Political	8	6	14
Infrastructure	12	0	12
Demographics	2	3	5
Temporal Context	0	3	3
TOTAL	45	66	111

Document Change Log

Version	Date	Changes
1.0	2025-12-11	Initial release
2.0	2026-01-07	Added macro-economic, ACLED Hybrid NLP, seasonal features
3.0	2026-01-15	Corrected ACLED drivers to 5, added IODA initial
4.0	2026-01-17	IOM date → 2015-01-31, Econ date → 2003-12-01, IODA finalized (2022-02-01), Dynamic World added, variables listed vertically

Generated: 2026-01-17 | *Pipeline Version: January 2026*