Contents

CEWP Data Source Audit (Updated) **1**
Spatio-Temporal Resolutions & Imputation Methods . . . 1
Executive Summary . . . 2
1. Environmental Data . . . 2
1.1 CHIRPS Precipitation . . . 2
1.2 ERA5-Land Climate . . . 2
1.3 MODIS Vegetation (NDVI) . . . 3
1.4 VIIRS Nighttime Lights . . . 3
1.5 JRC Global Surface Water . . . 3
2. Conflict & Event Data . . . 4
2.1 ACLED (Armed Conflict Location & Event Data) . . . 4
2.2 ACLED Hybrid NLP Features (NEW) . . . 4
2.3 GDELT (Global Database of Events, Language, and Tone) . . . 4
3. Socio-Political Data . . . 5
3.1 EPR (Ethnic Power Relations) . . . 5
3.2 IOM DTM (Displacement Tracking Matrix) . . . 5
3.3 FEWS NET IPC (Food Security Phases) . . . 6
4. Economic Data . . . 6
4.1 Macro-Economic Indicators (NEW) . . . 6
4.2 WFP/FEWS NET Food Prices (EXPANDED) . . . 7
5. Infrastructure & Geography (Static) . . . 7
6. Demographics . . . 8
6.1 WorldPop Population . . . 8
7. Temporal Context Features (NEW) . . . 9
8. Imputation Strategy . . . 9
8.1 Default Strategy . . . 9
8.2 Feature-Specific Overrides . . . 9
9. Structural Break Handling (NEW) . . . 10
10. Resolution Transformations . . . 10
10.1 Spatial Resolution Mapping . . . 10
10.2 Temporal Resolution Mapping . . . 10
11. Temporal Transformations . . . 11
12. Known Data Gaps & Issues (Updated) . . . 11
13. Feature Count Summary . . . 12
By Source Category . . . 12
Feature Registry Coverage . . . 12
Document Change Log . . . 12

CEWP Data Source Audit (Updated)

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: 2026-01-07

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. 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	7	100m – 11km, hourly/daily	H3-5, 14-day
Conflict & Events	3	Point, daily	H3-5, 14-day
Socio-Political	3	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

---

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.

1.1 CHIRPS Precipitation

Parameter	Value
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` → `chirps_precip_anomaly`
Imputation	Forward-fill (limit=4 steps / 56 days)
Coverage	1981–present

1.2 ERA5-Land Climate

Parameter	Value
Source	ECMWF/ERA5_LAND/HOURLY (via GEE)

Parameter	Value
Native Spatial	0.1° (~11km)
Native Temporal	Hourly
Variables	Temperature (2m), Dewpoint (2m), Soil Moisture (Layer 1)
Aggregation	H3-5 zonal mean, 14-day mean
Output Columns	`era5_temp_anomaly`, `era5_soil_moisture_anomaly`
Imputation	Forward-fill (limit=4 steps)
Coverage	1950–present (~5-day lag)

1.3 MODIS Vegetation (NDVI)

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

1.4 VIIRS Nighttime Lights

Parameter	Value
Source	NASA/VIIRS/002/VNP46A2 (via GEE)
Native Spatial	500m
Native Temporal	Daily
Aggregation	H3-5 zonal mean, 14-day mean
Output Column	`ntl_mean` → `nightlights_intensity`
Imputation	Forward-fill (limit=4 steps)
Coverage	2012–present
Structural Break	Pre-2012 data = NULL (sensor launch date)

1.5 JRC Global Surface Water

Parameter	Value
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
Output Columns	`water_coverage_lag1`, `water_presence_lag1`
Imputation	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)

Parameter	Value
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
Output Columns	`fatalities_14d_sum`, `fatalities_1m_lag`, `conflict_density_10km`, `protest_count_lag1`, `riot_count_lag1`, `regional_risk_score_lag1`
Imputation	Zero-fill for missing (no events = 0)
Coverage	1997–present (CAR: 2000+)
Volume	~7,500+ events in CAR

2.2 ACLED Hybrid NLP Features (NEW)

Parameter	Value
Source	ACLED notes field processed by `process_acled_hybrid.py`
Native Spatial	Event-level (inherits H3 from parent event)
Native Temporal	Event date
Method	Hybrid: 8 semantic themes (topic modeling) + 6 regex drivers
Output Columns	`theme_context_0` through `theme_context_7`, `driver_resource_cattle`, `driver_resource_mining`, `driver_econ_taxation`, `driver_civilian_abduct`, `driver_civilian_loot`, `driver_political_coup`
Imputation	Zero-fill for missing
Coverage	Full ACLED history

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

Parameter	Value
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
Output Columns	`events_3m_lag`, `gdelt_decay_30d`, `gdelt_avg_tone_decay_30d`
Imputation	Zero-fill for missing
Coverage	2015–present (v2)
Notes	Uses FIPS code "CT" for CAR

---

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)

Parameter	Value
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
Output Columns	`ethnic_group_count`, `epr_excluded_groups_count`, `epr_discriminated_groups_count`, `epr_status_mean`
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)
Notes	27 polygons in CAR; 282 group-year records
MISSING	`epr_horizontal_inequality` (Gini) - declared in config but no generation code

3.2 IOM DTM (Displacement Tracking Matrix)

Parameter	Value
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 Columns	`iom_displacement_count_lag1`, `iom_data_available` (structural break flag)
Imputation	Zero-fill between survey rounds
Coverage	2014–present
Notes	Fallback to Admin-1 if Admin-2 unavailable

3.3 FEWS NET IPC (Food Security Phases)

Parameter	Value
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	`ipc_phase_class` (1-5 scale)
Imputation	Constant=0 before 2009-01-01 (pre-IPC era); forward-fill after
Coverage	2009–present
Notes	Requires FEWS_NET_TOKEN in .env

---

4. Economic Data

Economic indicators include global commodities (non-spatial) and local market prices (market-specific).

4.1 Macro-Economic Indicators (NEW)

Parameter	Value
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)

Parameter	Value
Native Temporal	Daily (trading days)
Aggregation	National-level, 14-day mean
Output Columns	`gold_price_usd_lag1`, `oil_price_usd_lag1`, `sp500_index_lag1`, `eur_usd_rate_lag1`
Imputation	Forward-fill for non-trading days
Coverage	2000-08-30 (Gold futures) to present
Structural Break	`econ_data_available` flag for data availability

4.2 WFP/FEWS NET Food Prices (EXPANDED)

Parameter	Value
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
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` (12-month lookback)
Imputation	Forward-fill
Coverage	2015–present
Notes	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_index`	Via GEE

Data Source	Native Resolution	Output Features	Notes
GRIP4 Roads	Polyline network	`dist_to_road` (km)	PBL Region 3 (Africa)
HydroRIVERS	Polyline network	`dist_to_river` (km)	Stream order 3
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` (km)	914+ sites
OSM Settlements	Point locations	`dist_to_city`, `dist_to_capital` (km)	Via HDX
CAR Boundary	Polygon	`dist_to_border` (km)	World Bank boundaries

---

6. Demographics

6.1 WorldPop Population

Parameter	Value
Source	WorldPop 100m constrained (R2025A)
Native Spatial	100m raster
Native Temporal	Annual
Aggregation	H3-5 zonal sum, annual
Output Columns	`pop_count` → `pop_log` (log1p transformation), `is_worldpop_v1` (structural break flag)
Imputation	Forward-fill within year; backward extrapolation (2.5% annual) pre-2015
Coverage	2000–2030 (projections 2026-2030)
File Variants	`caf_ppp_{year}_UNadj.tif` (2000-2014), `caf_pop_{year}_CN_100m_R2025A_v1.tif` (2015+)
Structural Break	2015: V1 (census-adjusted) → V2 (constrained model)

---

7. Temporal Context Features (NEW)

Seasonal and cyclical features are generated from the date directly, not from external data sources.

Feature	Source	Description
`month_sin`	Temporal	Sine of month (captures cyclical seasonality)
`month_cos`	Temporal	Cosine of month (captures cyclical seasonality)
`is_dry_season`	Temporal	Binary flag: 1 if Nov-Mar, 0 otherwise

---

8. Imputation Strategy

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

8.1 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.

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

8.2 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
IOM displacement	Zero-fill	Between survey rounds
EPR features	Zero-fill	Cells outside ethnic polygons
VIIRS NTL	NULL	Pre-2012 (sensor not launched)
Economy	Zero-fill	Pre-2000-08-30 with `econ_data_available=0`

---

9. Structural Break Handling (NEW)

The pipeline now explicitly tracks data availability shifts that could confound model learning:

Break Flag	Affected Period	Purpose
`is_worldpop_v1`	Pre-2015	Distinguishes census-adjusted (V1) vs constrained (V2) population
`iom_data_available`	Pre-2014	IOM DTM data starts 2014
`econ_data_available`	Pre-2000-08-30	Yahoo Finance coverage start

These flags enable the model to learn different relationships for each methodological period rather than treating imputed values as true observations.

---

10. Resolution Transformations

10.1 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
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

10.2 Temporal Resolution Mapping

Native Frequency	Sources	Transformation	Method
Hourly	ERA5	14-day mean	GEE server-side
Daily	CHIRPS, ACLED, GDELT	14-day sum/mean	Client-side
16-day composite	MODIS	14-day max	Overlapping windows
Monthly	JRC water, WFP prices	14-day interpolation	Forward-fill within month

Native Frequency	Sources	Transformation	Method
Quarterly	IOM, IPC	14-day forward-fill	Limit=4 steps
Annual	WorldPop, EPR	Joined to all steps in year	Broadcast annual → 14-day
Static	DEM, roads, rivers, mines	Joined to all time steps	Cross join

---

11. 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)	`protest_count_lag1`
sum_1_step	Sum within period	`fatalities_14d_sum`
decay_30d	Exponential decay (half-life ~30 days, span=2.14)	`conflict_density_10km`
decay_90d	Exponential decay (half-life ~90 days, span=6.43)	`events_3m_lag`
anomaly	Value minus 6-period rolling mean	`chirps_precip_anomaly`
shock_12m	Value / 12-month rolling median	`price_maize_shock`
log1p	Natural log of (1 + x)	`pop_log`

---

12. Known Data Gaps & Issues (Updated)

Issue	Impact	Status	Priority
EPR horizontal_inequality missing	Feature in models.yaml but no generation code	Critical	HIGH
Seasonal features not in registry	month_sin, month_cos, is_dry_season undocumented	Documentation	MEDIUM
Distance features missing from features.yaml	All dist_* features generated but not in registry	Documentation	MEDIUM

Issue	Impact	Status	Priority
FEWS_NET_TOKEN missing	PC + market prices = NULL for some users	User setup	MEDIUM
Market locations CSV BOM	Encoding issue caused 0 markets loaded	Fixed	CLOSED
VIIRS pre-2012	NTL = NULL	Expected	CLOSED
EPR post-2021	Uses 2021 data for 2022-2025	Acceptable	CLOSED
IOM survey gaps	Irregular temporal coverage	Zero-filled	CLOSED
h3 API deprecation	h3.k_ring → h3.grid_disk	Migrated	CLOSED

---

13. Feature Count Summary

By Source Category

Category	Raw Columns	Transformed Features	Total
Environmental	8	16 (with anomalies, lags)	24
Conflict	5	15 (with decays, lags, regional)	20
Economic	8	12 (with lags, shocks)	20
Socio-Political	8	4 (with lags)	12
Infrastructure	12	0 (static)	12
Demographics	2	3 (with log, structural break)	5
Temporal Context	0	3 (seasonal)	3
Total	**43**	**53**	**96**

Feature Registry Coverage

Status	Count	Percentage
Fully documented in features.yaml	65	68%
Generated but not in registry	28	29%
Declared but not implemented	3	3%

Document Change Log

Version	Date	Changes
1.0	2025-12-11	Initial release

Version	Date	Changes
2.0	2026-01-07	**Major update:** Added macro-economic indicators (Section 4.1), Added ACLED Hybrid NLP features (Section 2.2), Added seasonal features (Section 7), Added structural break handling (Section 9), Updated imputation strategies, Added EPR missing feature documentation, Expanded feature counts to 96 total, Updated validation checkpoints

---

Generated: 2026-01-07
Auditor: CEWP Development Team
Pipeline Version: January 2026 (Phase 5 Complete)
Next Review: After next major feature addition or data source integration