Contents

Conflict Early Warning Pipeline (CEWP)

A Spatio-Temporal Machine Learning Framework for Predicting Armed Conflict Risk in the Central African Republic

Author: Brenan

Degree: Master's Thesis

Date: January 2026

Abstract

This thesis develops the Conflict Early Warning Pipeline (CEWP), an end-to-end geospatial data system that produces conflict-risk forecasts for the Central African Republic (CAR). The system integrates diverse data sources—satellite imagery, climate reanalysis, conflict event databases, economic indicators, and ethnic power relations—into a unified analytical framework using H3 hexagonal indexing at approximately 10km resolution.

A Two-Stage Hurdle Ensemble model architecture combines 9 thematic feature subsets (plus an optional PCA-based meta-theme) through XGBoost base learners and meta-learning aggregation, predicting conflict probability via Sigmoid-calibrated logistic regression and expected fatality counts via Poisson regression at 14-day, 1-month, and 3-month horizons. The pipeline is designed to support UN peacekeeping operations and humanitarian response planning by providing actionable early warning intelligence with calibrated uncertainty estimates.

Keywords: conflict prediction, early warning systems, machine learning, geospatial analysis, Central African Republic, humanitarian operations, ensemble learning, spatio-temporal modeling

1. Introduction

Armed conflict remains one of the most devastating challenges facing the international humanitarian community. In the Central African Republic, a country that has experienced recurring cycles of violence since 2012, the ability to anticipate where and when conflict will occur could fundamentally transform how humanitarian organizations allocate resources and protect civilian populations.

This thesis presents the Conflict Early Warning Pipeline (CEWP), a comprehensive data engineering and machine learning system designed to generate actionable conflict-risk forecasts. The system addresses three interconnected challenges: integrating heterogeneous data sources with varying spatio-temporal resolutions, engineering predictive features that capture the complex drivers of conflict, and producing probabilistic forecasts with meaningful uncertainty quantification.

1.1 Research Questions

ID	Research Question

RQ1	Can diverse geospatial data sources be systematically integrated into a unified analytical framework suitable for conflict prediction?
RQ2	Which thematic feature categories (environmental, economic, socio-political, conflict history) contribute most to predictive performance?
RQ3	How do prediction horizons (14-day, 1-month, 3-month) affect model performance and operational utility?
RQ4	Can ensemble methods improve upon single-model approaches for conflict forecasting?

1.2 Contributions

1. **Data Integration Framework:** A modular ETL pipeline that harmonizes 21+ distinct data sources to a common spatio-temporal grid, with documented imputation strategies for missing values.
2. **Feature Engineering Methodology:** A systematic approach to deriving conflict-relevant features from raw data, including temporal transformations (lags, decays, anomalies) and spatial aggregations.
3. **Two-Stage Hurdle Ensemble:** A novel model architecture that separately predicts conflict occurrence (classification) and intensity (regression), then combines thematic predictions through meta-learning.
4. **Operational Pipeline:** A production-ready system with incremental processing, caching, and error handling suitable for deployment in humanitarian contexts.

2. Background

2.1 The Central African Republic Context

The Central African Republic has experienced protracted instability since the 2012 Séléka rebellion and subsequent Anti-balaka reprisals. The conflict has displaced over 700,000 people internally and created conditions of chronic food insecurity affecting approximately half the population. The United Nations Multidimensional Integrated Stabilization Mission (MINUSCA) has operated in the country since 2014, representing one of the UN's largest peacekeeping deployments.

The conflict in CAR is characterized by several features that make it amenable to data-driven prediction: it is geographically diffuse rather than concentrated along a single front; it involves multiple armed groups with distinct territorial bases; and it exhibits seasonal patterns linked to agricultural cycles and road accessibility.

2.2 Conflict Prediction Literature

The field of conflict prediction has evolved substantially over the past two decades, moving from qualitative expert assessments toward quantitative forecasting systems. Key findings from the literature inform this thesis:

- Ensemble methods consistently outperform individual models in conflict forecasting competitions
- Spatial dependencies—through neighborhood features or graph neural networks—improve accuracy for geographically clustered events
- Model performance degrades predictably with longer forecast horizons, suggesting fundamental limits to predictability

2.3 Early Warning Systems in Practice

Operational early warning systems face constraints that academic forecasting models often ignore: data must be available in near-real-time, models must produce interpretable outputs, and forecasts must be specific enough to inform resource allocation decisions. The CEWP system is designed with these operational requirements in mind.

3. Methodology

3.1 Spatial Framework

The system uses Uber's H3 hierarchical hexagonal grid at resolution 5, which produces cells of approximately 253 km² (roughly 10km edge-to-edge).

Parameter	Value	Rationale
Grid System	H3 Hexagonal	Uniform adjacency, no orientation bias
Resolution	5	~253 km² cells, ~10km diameter
Coverage	3,407 cells	Full CAR territory
CRS (Analysis)	EPSG:32634	UTM Zone 34N for metric operations
CRS (Storage)	EPSG:4326	WGS84 for interoperability

3.2 Temporal Framework

Parameter	Value

Base Unit	14-day steps
Epoch	2000-01-01
Analysis Period	2000–2025 (26 years)
Prediction Horizons	14 days (1 step), 1 month (2 steps), 3 months (6 steps)
Lag Features	1, 2, 3 steps (14, 28, 42 days)
Decay Half-lives	30 days (~2.14 steps), 90 days (~6.43 steps)

3.3 Data Integration Pipeline

The pipeline integrates 21+ distinct data sources across eight thematic categories:

Category	Sources	Key Features
Environmental	CHIRPS, ERA5, MODIS, VIIRS, JRC Water, Dynamic World	Precipitation, temperature, NDVI, nighttime lights, surface water, landcover
Conflict	ACLED, GDELT	Event counts, fatalities, protest/riot indicators, media tone
ACLED NLP	ACLED notes field	8 semantic themes + 5 explicit drivers
Socio-Political	EPR, IOM DTM, FEWS NET IPC, IODA	Ethnic exclusion, displacement, food security, internet outage events
Economic	Yahoo Finance, WFP Markets	Commodity prices (gold, oil), local market prices, price shocks
Infrastructure	GRIP4, HydroRIVERS, IPIS, OSM	Distance to roads, rivers, mines, settlements
Demographics	WorldPop	Population count and density
Temporal	Generated	Seasonal features (month_sin, month_cos, is_dry_season)

3.4 Feature Engineering

3.4.1 Temporal Transformations

Transform	Formula	Purpose
Lag	x(t-k)	Capture delayed effects; prevent leakage
Sum	Σ x(t-k:t)	Accumulate events within window
Decay	EWM(x, span=λ)	Weight recent events more heavily
Anomaly	x(t) - μ(t-6:t)	Deviation from rolling baseline
Shock	x(t) / median(t-12:t)	Price spikes relative to historical norm

3.4.2 Temporal Lag Handling

The pipeline distinguishes between two independent lag mechanisms:

Publication Lags (Data Availability): Account for the delay between data collection and public availability. Applied at ingestion/storage so timestamps reflect when data would actually be available. Examples: GEE +14 days, Food Prices +56 days, ACLED NLP +14 days.

Analytical Lags (Leakage Control): Ensure features only use prior-period values, preventing temporal leakage. Applied downstream via LAG()/shift() for features and LEAD() for targets.

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.

3.4.3 Imputation Strategy

Feature Type	Method	Details
Default	Forward-fill	limit=4 steps (56 days)
Conflict events	Zero-fill	No event = 0
Population	Forward-fill + zero	Forward-fill within hex; pre-coverage gaps = 0 (no backward extrapolation)
IPC Phase	Constant	Value=0 before 2009

3.4.4 Structural Break Handling

Flag	Period	Purpose
is_worldpop_v1	Pre-2015	V1 census-adjusted vs V2 constrained
iom_data_available	Pre-2015-01-31	IOM DTM coverage start
econ_data_available	Pre-2003-12-01	Yahoo Finance coverage start
ioda_data_available	Pre-2022-02-01	IODA internet monitoring start
landcover_avail	Pre-2015-06-27	Dynamic World landcover start

4. Model Architecture

4.1 Two-Stage Hurdle Ensemble

The CEWP employs a Two-Stage Hurdle Ensemble architecture designed to address the dual prediction tasks of conflict occurrence (binary) and conflict intensity (count).

Stage 1: 9 thematic sub-models (+ optional Broad PCA), each operating on a distinct feature subset

Stage 2: Logistic and Poisson meta-learners aggregate predictions, enforcing non-negativity and discrete count constraints

Stage 3: Sigmoid (Platt) calibration maps raw scores to calibrated probabilities, preventing overfitting on rare event classes

Stage 4: BCCP provides prediction intervals with guaranteed coverage

4.2 Thematic Sub-Models (9 + Broad PCA)

Theme	Features	Hypothesis
Baseline	fatalities_1m_lag, pop_log	Persistence and exposure
Conflict History	ACLED structured: fatalities (decay), protests, riots, regional_risk_score	Escalation dynamics from curated event categories
News & Ops	GDELT tone/counts, CrisisWatch alerts, IODA outages, ACLED hybrid drivers	Abstract signals: sentiment, language,

		connectivity
Environmental	Precip/temp/NDVI anomalies, nightlights, water, landcover	Climate stress
Terrain	Elevation, slope, TRI, distances to infrastructure	Accessibility and refuge
Economics	Gold price, oil price, food prices, price shocks	Economic grievances
EPR	Excluded groups, status mean	Ethnic power dynamics
Demographic	pop_log, is_worldpop_v1	Population exposure with structural break
Temporal Context	month_sin, month_cos, is_dry_season	Seasonal patterns
Broad PCA (Optional)	PCA components from all enabled themes (90% variance)	Cross-feature structure and dimensionality reduction

Note: Conflict History uses ACLED's curated, structured event categories (event types, actor codes, geo-precision). News & Ops carries more abstract variables—GDELT media sentiment/tone, CrisisWatch alerts, IODA connectivity outages, and ACLED Hybrid NLP drivers (semantic themes extracted from free-text notes).

4.2.1 Broad PCA Theme

The Broad PCA theme is an optional meta-feature generator that captures cross-feature structure across all other enabled themes:

- During training, process_pca_subsampled() collects all features from enabled submodels (excluding broad_pca itself), samples up to 300k rows, scales them, and fits PCA to retain 90% variance
- The fitted PCA transforms the full dataset in chunks, appending columns pca_1...pca_k to the feature matrix
- build_theme_models() then treats broad_pca as just another theme, using the PCA component columns as inputs
- At inference, the saved PCA object and scaler are loaded to recreate the same pca_* columns before prediction

Net effect: an additional theme in the ensemble stack that captures latent cross-feature relationships. It only runs if enabled in configuration and the PCA fit succeeds; otherwise it is skipped gracefully.

4.3 Training Procedure

- Temporal cross-validation with expanding windows
- Training data: All observations through December 2020
- Test data: 2021–2025 (validation: 2021, 2022, 2024, 2025; holdout test: 2023)
- TimeSeriesSplit: 5 folds for out-of-fold prediction generation
- Class imbalance: Dynamic scale_pos_weight (default ~35.0 for 2% positive class)

5. Evaluation Framework

Metric	Stage	Interpretation
PR-AUC	Stage 1 (Occurrence)	Discrimination for imbalanced data
ROC-AUC	Stage 1 (Occurrence)	Overall classification performance
Brier Score	Stage 1 (Occurrence)	Calibration quality (lower is better)
Top-10% Recall	Stage 1 (Occurrence)	Operational efficiency
Mean Poisson Deviance	Stage 2 (Intensity)	Count model fit (lower is better)
RMSE	Stage 2 (Intensity)	Intensity prediction accuracy (interpretability)

Mean Poisson Deviance (D) measures the goodness-of-fit for count data models. Unlike RMSE, which assumes constant variance (homoscedasticity), MPD accounts for the heteroscedastic nature of conflict data, where variance increases with the mean. A lower deviance indicates that the model is better at capturing the relative magnitude of conflict events.

Operational Focus: The Top-10% Recall metric is particularly important—if a humanitarian organization can only monitor 10% of the geographic area, this metric indicates what fraction of actual conflicts would fall within their coverage.

6. System Design

6.1 Pipeline Architecture

Phase	Components	Outputs

1. Static Ingestion	H3 grid, DEM, roads, rivers, mines, settlements, EPR	features_static table
2. Dynamic Ingestion	ACLED, GDELT, IOM, GEE environmental, Economy, IODA	features_dynamic_daily, environmental_features
3. ACLED NLP	Semantic topics, regex drivers	features_acled_hybrid table
4. Feature Engineering	Temporal transforms, spatial aggregation, imputation	temporal_features table
5. Model Training	Theme models, meta-learners, cross-validation	Serialized model artifacts
6. Inference	Prediction generation, uncertainty quantification	Risk forecasts (Parquet/GeoJSON)

6.2 Technical Stack

Component	Technology
Database	PostgreSQL 15 with PostGIS 3.4
Spatial Indexing	H3 (Uber), GEOS
Raster Processing	Rasterio, GDAL
ML Framework	XGBoost, LightGBM, Scikit-learn
Cloud APIs	Google Earth Engine, BigQuery
Configuration	YAML (data.yaml, features.yaml, models.yaml)

6.3 Configuration-Driven Design

All features, data sources, and model parameters are defined in YAML configuration files, enabling:

- Reproducible experiments
- Easy feature addition/removal
- Clear lineage from config → code → database

7. Results Summary

(To be completed with actual evaluation metrics)

7.1 Model Performance by Horizon

Horizon	PR-AUC	Brier Score	Top-10% Recall	Mean Poisson Deviance
14-day	-	-	-	-
1-month	-	-	-	-
3-month	-	-	-	-

7.2 Feature Importance (Preliminary)

Top features by SHAP importance:

1. fatalities_1m_lag - Recent conflict persistence
2. regional_risk_score_lag1 - Administrative-level spillover
3. conflict_density_10km - Spatial clustering
4. pop_log - Population exposure
5. Seasonal features (month_sin, is_dry_season)

8. Discussion

8.1 Limitations

1. **Data quality:** ACLED's event coverage depends on media reporting and may undercount violence in remote areas
2. **Temporal resolution:** 14-day windows may miss rapid escalation dynamics
3. **Causal mechanisms:** Current features capture correlates, not causes, limiting policy interpretability

8.2 Ethical Considerations

- Forecasts could be misused for preemptive military action rather than humanitarian protection
- Risk maps might stigmatize high-risk communities
- Responsible deployment requires ongoing engagement with local stakeholders

8.3 Future Directions

1. **Graph Neural Networks:** Explicit spatial dependencies through ST-GNN architectures
2. **Uncertainty Quantification:** Bin-Conditional Conformal Prediction (BCCP) for prediction intervals
3. **Causal Discovery:** Identify actionable intervention points
4. **Real-time Deployment:** Streaming data pipelines for near-real-time forecast updates

9. Conclusion

This thesis has presented the Conflict Early Warning Pipeline, a comprehensive system for generating conflict-risk forecasts in the Central African Republic. The pipeline demonstrates that:

1. Diverse geospatial data sources can be systematically integrated into a unified analytical framework
2. Thematic feature engineering captures interpretable conflict drivers
3. Ensemble methods improve upon single-model approaches for rare-event prediction
4. Structural break handling enables learning across methodological changes in source data

The modular pipeline design supports operational deployment with incremental updates, robust error handling, and configurable parameters. By providing timely, spatially explicit risk estimates with calibrated uncertainty, the system aims to enable more proactive and effective humanitarian response in one of the world's most protracted crises.

Appendix A: Complete Feature Registry

Total Features: 111 (45 raw + 66 transformed)

A.1 Environmental Features (26 total)

Feature	Source	Transform
chirps_precip_anomaly	CHIRPS	anomaly_6_step
era5_temp_anomaly	ERA5	anomaly_6_step
era5_soil_moisture_anomaly	ERA5	anomaly_6_step
ndvi_anomaly	MODIS	anomaly_6_step
nightlights_intensity	VIIRS	mean
water_coverage_lag1	JRC/Landsat	lag_1_step

water_presence_lag1	JRC/Landsat	lag_1_step
landcover_grass	Dynamic World	mean fraction (0-1)
landcover_crops	Dynamic World	mean fraction (0-1)
landcover_trees	Dynamic World	mean fraction (0-1)
landcover_bare	Dynamic World	mean fraction (0-1)
landcover_built	Dynamic World	mean fraction (0-1)

A.2 Conflict Features (20 total)

Feature	Source	Transform
fatalities_14d_sum	ACLED	sum_1_step
fatalities_1m_lag	ACLED	lag
conflict_density_10km	ACLED	decay_30d
protest_count_lag1	ACLED	lag_1_step
riot_count_lag1	ACLED	lag_1_step
regional_risk_score_lag1	ACLED	lag_1_step
events_3m_lag	GDELT	decay_90d
gdelt_decay_30d	GDELT	decay_30d
gdelt_avg_tone_decay_30d	GDELT	decay_30d

A.3 ACLED Hybrid Features (13 total)

Feature	Description

theme_context_0 - theme_context_7	Semantic topic weights from event notes (8 themes)
driver_resource_cattle	Cattle-related conflict indicator
driver_resource_mining	Mining-related conflict indicator
driver_econ_taxation	Taxation/economic conflict indicator
driver_political_coup	Coup/political violence indicator
driver_civilian_abuse	Human rights violations indicator

A.4 Economic Features (20 total)

Feature	Source	Transform
gold_price_usd_lag1	Yahoo Finance	lag_1_step
oil_price_usd_lag1	Yahoo Finance	lag_1_step
sp500_index_lag1	Yahoo Finance	lag_1_step
eur_usd_rate_lag1	Yahoo Finance	lag_1_step
price_maize	FEWS NET	none
price_maize_shock	FEWS NET	shock_12m
price_rice / price_rice_shock	FEWS NET	none / shock_12m
price_oil / price_oil_shock	FEWS NET	none / shock_12m
price_sorghum / price_sorghum_shock	FEWS NET	none / shock_12m
food_price_index	FEWS NET	none
econ_data_available	Structural	none

A.5 Socio-Political Features (14 total)

Feature	Source	Transform
epr_excluded_groups_count	EPR	none
epr_discriminated_groups_count	EPR	none
epr_status_mean	EPR	none
ethnic_group_count	EPR	none
iom_displacement_count_lag1	IOM DTM	lag_1_step
iom_data_available	Structural	none
ipc_phase_class	FEWS NET IPC	none
ioda_outage_score	IODA	none
ioda_data_available	Structural	none

A.6 Infrastructure Features (13 total)

Feature	Source
dist_to_capital	OSM
dist_to_border	CAR Boundary
dist_to_road	GRIP4
dist_to_city	OSM
dist_to_river	HydroRIVERS
dist_to_diamond_mine	IPIS
dist_to_gold_mine	IPIS

dist_to_large_mine	IPIS
dist_to_controlled_mine	IPIS
dist_to_large_gold_mine	IPIS
terrain_ruggedness_index	Copernicus DEM
elevation_mean	Copernicus DEM
slope_mean	Copernicus DEM

A.7 Demographic Features (5 total)

Feature	Source	Transform
pop_count	WorldPop	none
pop_log	WorldPop	log1p
is_worldpop_v1	Structural	none

A.8 Temporal Context Features (3 total)

Feature	Description
month_sin	Sine transformation of month
month_cos	Cosine transformation of month
is_dry_season	Binary: 1 if Nov-Mar

Appendix B: Database Schema

car_cewp schema:

features_static (spatial foundation)

Columns: h3_index (BIGINT PK), geometry, elevation_mean, slope_mean, terrain_ruggedness_index, dist_to_*, admin1, admin2, admin3

temporal_features (time series)

Columns: h3_index, date (composite PK), [environmental], [conflict], [economic], [socio-political]

features_acled_hybrid (NLP features)

Columns: h3_index, date (composite PK), theme_context_0-7, driver_resource_*, driver_econ_*, driver_civilian_*, driver_political_*

Raw ingestion tables:

- acled_events (h3_index, event_date, fatalities, event_type, ...)
- environmental_features (h3_index, date, precip_mean_depth_mm, ...)
- economic_drivers (date, gold_price_usd, oil_price_usd, ...)
- food_security (date, market, commodity, value)
- iom_displacement_h3 (h3_index, date, iom_displacement_sum)
- ioda_outages (h3_index, date, outage_score)

Spatial reference tables:

- population_h3 (h3_index, year, pop_count)
- grip4_roads_h3 (h3_index, road_density)
- geoepr_polygons (group_id, status, geometry)
- market_locations (market_id, latitude, longitude)

Appendix C: System Requirements

Hardware

- RAM: 16GB minimum (32GB recommended for full pipeline)
- Storage: 50GB for raw data + 20GB for processed outputs
- CPU: Multi-core recommended for parallel processing

Software

- Python 3.10+
- PostgreSQL 13+ with PostGIS 3.0+ and H3 extension
- Google Earth Engine account (for satellite data)

Document Version: 2.1
Last Updated: January 2026
Total Features: 111
Thematic Sub-models: 9 + Broad PCA