Conflict Early Warning Pipeline

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

Brenan

Master's Thesis

December 2025

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 thematic feature subsets through XGBoost base learners and meta-learning aggregation, predicting both the probability of conflict occurrence and expected fatality counts 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

This research is guided by the following questions:

	Research Question
RQ1	Can diverse geospatial data sources be systematically integrated into a unified analytical framework suitable for conflict p
RQ2	Which thematic feature categories (environmental, economic, socio-political, conflict history) contribute most to predictive
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?

Table 1: Research questions guiding this thesis

1.2 Contributions

This thesis makes the following contributions to the fields of conflict forecasting and humanitarian data science:

- **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.
- **Feature Engineering Methodology:** A systematic approach to deriving conflict-relevant features from raw data, including temporal transformations (lags, decays, anomalies) and spatial aggregations (k-ring neighbors).
- **Two-Stage Hurdle Ensemble:** A novel model architecture that separately predicts conflict occurrence (classification) and intensity (regression), then combines thematic predictions through meta-learning.
- **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. These characteristics suggest that environmental, economic, and political indicators may provide meaningful predictive signals.

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. Early work focused on country-level risk indices, while more recent approaches have pursued sub-national prediction at increasingly fine spatial and temporal resolutions.

Several key findings from the literature inform this thesis. First, ensemble methods consistently outperform individual models in conflict forecasting competitions. Second, the inclusion of spatial dependencies—through neighborhood features or graph neural networks—improves accuracy for geographically clustered events. Third, model performance degrades predictably with longer forecast horizons, suggesting a fundamental limit to predictability in complex social systems.

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, not just historically. Models must produce outputs interpretable by non-technical decision-makers. Forecasts must be actionable—specific enough in space and time to inform resource allocation decisions. The CEWP system is designed with these operational requirements in mind.

3. Methodology

The CEWP methodology consists of four integrated components: spatial representation, data integration, feature engineering, and predictive modeling. Each component is described in detail below.

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). This resolution balances spatial granularity against data sparsity—finer resolutions would create cells with insufficient conflict events for reliable estimation, while coarser resolutions would obscure meaningful spatial variation.

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

Table 2: Spatial framework parameters

3.2 Temporal Framework

All time-varying data is aggregated to 14-day windows aligned to a fixed epoch (January 1, 2000). This alignment ensures consistent temporal joins across data sources and creates a regular time series suitable for lag-based feature engineering. The 14-day window represents a compromise between temporal resolution and data density—shorter windows would increase sparsity in conflict events, while longer windows would reduce the system's responsiveness to rapid changes.

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 six thematic categories. Each source undergoes extraction, transformation, and loading (ETL) to produce features aligned to the common spatio-temporal grid. The integration process handles three classes of spatial data: rasters (aggregated via zonal statistics), points (assigned via point-in-polygon), and administrative polygons (disaggregated via area-weighted distribution or polygon overlay).

Category	Sources	Key Features
Environmental	CHIRPS, ERA5, MODIS, VIIRS, JRC Water	Precipitation, temperature, NDVI, nighttime lights, surface water
Conflict	ACLED, GDELT	Event counts, fatalities, protest/riot indicators, media tone
Socio-Political	EPR, IOM DTM, FEWS NET IPC	Ethnic exclusion, displacement, food security phase
Economic	Yahoo Finance, WFP Markets	Commodity prices, local market prices
Infrastructure	GRIP4, HydroRIVERS, IPIS, OSM	Distance to roads, rivers, mines, settlements
Demographics	WorldPop	Population count and density

Table 3: Data sources by thematic category

3.4 Feature Engineering

Raw data undergoes systematic transformation to create predictive features. The feature engineering process applies temporal transformations to capture dynamics and spatial transformations to capture neighborhood effects.

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=α)	Weight recent events more heavily
Anomaly	x(t) - μ(t-6:t)	Deviation from rolling baseline

Table 4: Temporal transformation functions

3.4.2 Spatial Transformations

Spatial features capture neighborhood context through H3 k-ring operations. For each cell, we compute aggregate statistics (mean, max, sum) over neighboring cells at distances k=1, 2, and 3. This approach captures spatial spillover effects—conflict in neighboring areas may predict escalation in the focal cell.

3.4.3 Imputation Strategy

Missing values are handled through feature-specific strategies that preserve temporal coherence and avoid information leakage. The default strategy is forward-fill with a 4-step (56-day) limit. Event-based features (conflict, displacement) are zero-filled under the assumption that missing data indicates no observed events. Population data undergoes backward extrapolation with a 2.5% annual growth rate for years prior to 2015 when high-resolution estimates begin.

3.5 Model Architecture

The CEWP employs a Two-Stage Hurdle Ensemble architecture designed to address the dual prediction tasks of conflict occurrence (binary) and conflict intensity (count). This architecture reflects the empirical reality that the factors predicting whether conflict occurs may differ from those predicting severity conditional on occurrence.

3.5.1 Stage 1: Thematic Base Learners

The first stage consists of five thematic sub-models, each operating on a distinct feature subset. This design serves multiple purposes: it reduces overfitting by limiting each model's feature space, enables interpretable attribution of predictive power to thematic categories, and allows the ensemble to learn complementary patterns across themes.

Theme	Features	Hypothesis
Baseline	Fatalities (14d sum), Population (log)	Persistence and exposure
Conflict History	Fatalities (decay), Protests, Riots	Escalation dynamics
Geography	Distance to capital/roads, Terrain, Water	Accessibility and refuge
Resources & Economy	Mining distance, Gold price, Nighttime lights	Economic grievances
Socio-Politics	Ethnic exclusion, Displacement, Food security	Structural vulnerability

Table 5: Thematic feature subsets and theoretical motivation

Each theme contains two XGBoost models: a classifier predicting P(conflict > 0) and a regressor predicting E[fatalities | conflict > 0]. The classifier is trained on all observations; the regressor is trained only on observations where conflict occurred, following the hurdle model specification.

3.5.2 Stage 2: Meta-Learning

The second stage aggregates predictions from the thematic models through meta-learners. Out-of-fold predictions from Stage 1 are used as features for Stage 2, preventing information leakage. A Logistic Regression meta-classifier combines the five binary probability estimates; a Ridge Regression meta-regressor combines the five fatality predictions. The use of linear meta-learners constrains complexity and promotes interpretable ensemble weights.

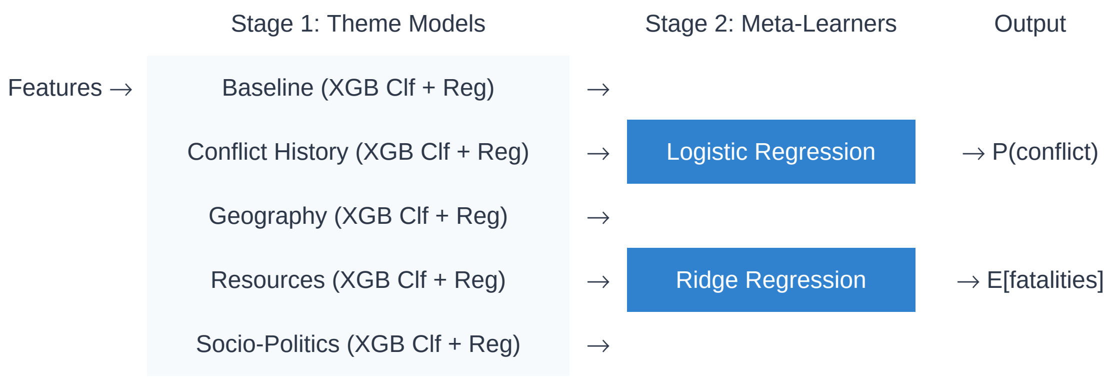
Stage 1: Theme Models
Stage 2: Meta-Learners
Output
Features →
Baseline (XGB Clf + Reg) →
Conflict History (XGB Clf + Reg) →
Geography (XGB Clf + Reg) →
Resources (XGB Clf + Reg) →
Socio-Politics (XGB Clf + Reg) →
Logistic Regression
→ P(conflict)
Ridge Regression
→ E[fatalities]

Figure 1: Two-Stage Hurdle Ensemble architecture

3.5.3 Training Procedure

Models are trained using temporal cross-validation with expanding windows to respect the time-series nature of the data. Training data includes all observations through December 2020; test data spans 2021–2025. Within the training period, TimeSeriesSplit generates 5 folds for out-of-fold prediction generation. The classifier training employs optional downsampling to address class imbalance (conflict events are rare), while validation folds are never downsampled to ensure unbiased performance estimates.

4. Evaluation Framework

Model performance is assessed using metrics appropriate for rare-event prediction and operational decision-making. The evaluation framework emphasizes discrimination (can the model rank observations by risk?), calibration (are probability estimates reliable?), and operational utility (does the model identify high-risk areas efficiently?).

Metric	Formula / Description	Interpretation
PR-AUC	Area under Precision-Recall curve	Discrimination for imbalanced data
Brier Score	Mean squared error of probabilities	Calibration quality (lower is better)
Top-10% Recall	True positives in highest risk decile / Total positives	Operational efficiency
RMSE	Root mean squared error of fatality predictions	Intensity prediction accuracy

Table 6: Evaluation metrics

The Top-10% Recall metric is particularly important for operational use. If a humanitarian organization can only monitor or respond to 10% of the geographic area, this metric indicates what fraction of actual conflicts would fall within their coverage. A random model would achieve 10% recall; values substantially higher indicate meaningful predictive power.

5. System Design

The CEWP is implemented as a modular Python pipeline with PostgreSQL/PostGIS for spatial data storage. The system architecture prioritizes reproducibility, incremental processing, and operational reliability.

5.1 Pipeline Architecture

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

Table 7: Pipeline phases and outputs

5.2 Technical Stack

Component	Technology

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

6. Discussion

6.1 Limitations

Several limitations should be acknowledged. First, the system relies on secondary data sources that may contain systematic biases—ACLED's event coverage, for example, depends on media reporting and may undercount violence in remote areas. Second, the 14-day temporal resolution, while appropriate for operational planning, may miss rapid escalation dynamics. Third, the current feature set captures correlates of conflict but does not model causal mechanisms, limiting interpretability for policy intervention.

6.2 Ethical Considerations

Conflict prediction systems raise significant ethical concerns. Forecasts could be misused for preemptive military action rather than humanitarian protection. Publication of risk maps might stigmatize high-risk communities or create self-fulfilling prophecies. The system is designed for humanitarian applications with appropriate access controls; responsible deployment requires ongoing engagement with local stakeholders and conflict sensitivity training for end users.

6.3 Future Directions

- **Graph Neural Networks:** Incorporate explicit spatial dependencies through Spatio-Temporal GNN architectures operating on the H3 adjacency graph.
- **Uncertainty Quantification:** Implement Bayesian Calibrated Conformal Prediction (BCCP) to produce prediction intervals with coverage guarantees.
- **Causal Discovery:** Apply causal inference methods to identify actionable intervention points within the feature space.
- **Real-time Deployment:** Develop streaming data pipelines for near-real-time forecast updates as new conflict events are reported.

7. 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 diverse geospatial data sources can be systematically integrated into a unified analytical framework, that thematic feature engineering captures interpretable conflict drivers, and that ensemble methods improve upon single-model approaches for rare-event prediction.

The Two-Stage Hurdle Ensemble architecture addresses the dual challenges of predicting conflict occurrence and intensity, while the meta-learning stage enables the model to learn optimal combinations of thematic predictions. The modular pipeline design supports operational deployment with incremental updates, robust error handling, and configurable parameters.

Ultimately, the value of conflict early warning systems lies not in their technical sophistication but in their contribution to protecting civilian lives. The CEWP is designed to support—not replace—human judgment in humanitarian decision-making. 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: Feature Registry

The following table documents all engineered features used in the model, their source data, and transformations applied.

Feature	Source	Transform
chirps_precip_anomaly	CHIRPS	anomaly_6_step
era5_temp_mean_lag1	ERA5	lag_1_step
ndvi_max_lag1	MODIS	lag_1_step
viirs_ntl_mean_lag1	VIIRS	lag_1_step
era5_soil_moisture_anomaly	ERA5	anomaly_6_step
water_coverage_lag1	JRC/Landsat	lag_1_step
fatalities_decay_30d	ACLED	decay_30d
fatalities_14d_sum	ACLED	sum_1_step
protest_count_lag1	ACLED	lag_1_step
riot_count_lag1	ACLED	lag_1_step
commodity_gold_price_usd_lag1	Yahoo Finance	lag_1_step
ipc_phase_class_lag1	FEWS NET	lag_1_step
iom_displacement_count_lag1	IOM DTM	lag_1_step

Table A1: Partial feature registry