# Probabilistic Machine Learning for Climate-Sensitive Cholera Outbreak Risk Prediction in Nigeria


## Abstract

Cholera remains a recurring public-health threat in Nigeria, where seasonal climate patterns, structural vulnerability, and persistent WASH deficits contribute to repeated sub-national outbreaks. This study develops a probabilistic early-warning framework for cholera outbreak risk using a state-month analytical panel covering Nigeria's 36 states and the Federal Capital Territory from 2018 to 2025. The panel integrates cholera surveillance data with rainfall, temperature, humidity proxy, WASH indicators, poverty, population, surface-water features, and engineered outbreak-history variables. Seven supervised models were evaluated using rolling time-aware cross-validation, including Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, a weighted probability blend, and a logistic meta-stacking ensemble. The overall outbreak rate across 1,591 panel rows was 26.8%. CatBoost achieved the best balanced performance (F1 = 0.556, ROC-AUC = 0.668), while Random Forest achieved the highest ROC-AUC (0.677) and Logistic Regression the best average precision (0.611). Outbreak-history features contributed the largest predictive signal, outperforming climate, WASH, and vulnerability features in isolation. Cholera burden was concentrated in the North-West and North-East, with peak seasonal risk in July, August, and September. The resulting system is best interpreted as a high-sensitivity preparedness support tool rather than a high-precision classifier, providing a practical basis for geographically targeted cholera response in Nigeria.


## 1. Introduction

holera remains a major public-health threat in sub-Saharan Africa, with Nigeria among the most consistently affected countries in the region. Outbreaks recur across a geographically concentrated subset of states and typically intensify during the rainy season, reflecting the interaction of environmental exposure, weak water and sanitation systems, and underlying structural vulnerability [WHO, 2023; NCDC, 2022]. Despite these recurrent seasonal and spatial patterns, preparedness in many settings remains more reactive than anticipatory, limiting the ability of health systems to act before transmission accelerates.

Effective early warning requires more than routine surveillance alone. Cholera risk is shaped by a layered interaction of climatic conditions, water and sanitation access, poverty, population concentration, surface-water exposure, and the persistence of prior outbreaks [Jutla et al., 2013; Ali et al., 2015]. This makes cholera a suitable target for probabilistic risk modeling, especially in resource-constrained settings where timely preparedness decisions such as pre-positioning supplies, intensifying surveillance, and targeting WASH response can substantially affect outbreak control.

Machine-learning methods offer useful advantages for this task because they can accommodate nonlinear relationships, threshold effects, and interactions across heterogeneous feature sets [Brownstein et al., 2008; Colwell et al., 2020]. However, many existing cholera prediction studies focus on a limited set of environmental drivers,

operate at coarse spatial resolution, or omit temporal outbreak-history features that may capture recurrent state-level vulnerability. In addition, model performance is often reported without sufficiently emphasizing the difference between experimental discrimination and practical deployment readiness.

This study addresses these limitations by developing a reproducible state-month predictive pipeline for cholera outbreak risk in Nigeria. Using an analytical panel covering the 36 states and the Federal Capital Territory from 2018 to 2025, the study integrates cholera surveillance data with rainfall, temperature, humidity proxy, WASH indicators, poverty, population, surface-water features, and engineered outbreak-history variables. Multiple model families are evaluated under rolling time-aware validation, and predicted probabilities are translated into interpretable alert categories for operational use. The paper makes four contributions: it constructs an integrated state-month cholera panel for Nigeria spanning 2018 to 2025; it shows that outbreak-history features contribute more predictive signal than climate variables in isolation; it provides a cross-model comparison under temporally honest evaluation; and it demonstrates a practical high-sensitivity alert framework for preparedness-oriented decision support.

## 2. Related Work

Cholera's environmental and climatic determinants have been widely studied. Rainfall, temperature, and surface-water availability influence the ecology and spread of *Vibrio cholerae*, while WASH access, poverty, and settlement conditions shape human vulnerability to transmission [Jutla et al., 2013; Mukandavire et al., 2011]. Seasonal cholera patterns are well documented across sub-Saharan Africa, where transmission often peaks during or shortly after the rainy season [Ali et al., 2015].

Predictive modeling efforts have increasingly incorporated remote sensing, climate reanalysis, and machine-learning approaches. Early work emphasized environmental suitability models based on rainfall and temperature anomalies [Bertuzzo et al., 2010], while later studies incorporated WASH and socioeconomic variables to improve sub-national risk estimation [Finger et al., 2014; Moore et al., 2017]. In Nigeria, prior spatial analyses have shown strong regional heterogeneity, with northern states carrying a disproportionate share of outbreak burden [Bolu et al., 2019].

Despite this progress, several limitations remain. Many prior models do not explicitly incorporate temporal outbreak-history features, which may capture endemic persistence in recurrent high-risk states. Lagged climate indicators have been explored in some cholera settings [Rinaldo et al., 2012], but their marginal contribution relative to surveillance history has not been systematically evaluated. Ensemble and gradient-boosted models have also been applied only sparingly in Nigerian cholera prediction under rolling time-aware validation. This study addresses these gaps within a unified state-month analytical pipeline.

## 3. Methods

### 3.1 Study Design and Data

The study used a repeated cross-sectional state-month panel design covering Nigeria's 36 states and the Federal Capital Territory from January 2018 to early 2025. The unit of analysis was the state-month. The analytical panel contained 1,591 rows across 37 administrative units, 43 observed month-periods, and a total of 41,898 compiled cholera cases. The overall outbreak rate was 0.268, with 426 outbreak months recorded. Forecast-ready rows after temporal feature construction numbered

1,369, with 296 held-out test rows including 110 positive outbreak months.

Data sources included: NCDC monthly cholera surveillance reports for outbreak labels and case burden; CHIRPS for rainfall totals, anomalies, and lagged rainfall; ERA5 for temperature and humidity proxy variables; WHO/UNICEF JMP for improved water and sanitation access indicators; WorldPop for population and population density; state-level poverty rate estimates with explicit missingness handling; and JRC Global Surface Water for occurrence, seasonality, and maximum extent metrics.

### 3.2 Outcome

The binary outcome was one-month-ahead cholera outbreak status, defined as any confirmed outbreak in a given state-month. This forward-looking specification ensures the model generates actionable early warning estimates rather than contemporaneous descriptive risk scores.

### 3.3 Feature Engineering

Temporal and epidemiological features were constructed to capture outbreak persistence and recent burden: lagged outbreak flags, lagged cholera case counts, rolling 3-month and 6-month outbreak sums, rolling 3-month and 6-month case burden, prior state outbreak rate, lagged case rates per 100,000, outbreak streak indicators, and month sine and cosine seasonality terms.

Interaction and derived features included: rainfall-poverty interaction, humidity-surface water interaction, rainfall-surface water interaction, temperature-rainfall interaction, a composite poor WASH index, temperature range, absolute rainfall anomaly, recent outbreak pressure, recent case pressure, lagged case momentum, and a seasonal surface-water interaction term.

### 3.4 Modeling Pipeline

Seven supervised models were compared: Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, a weighted probability blend, and a logistic meta-stacking ensemble. Preprocessing steps included rolling time-aware cross-validation to prevent data leakage; constrained hyperparameter search across model families; threshold calibration using F1.5-oriented selection; missing-value imputation; one-hot encoding for state and region; robust scaling on the logistic regression path; leakage-safe quantile clipping for numeric outliers; class weighting; and controlled majority undersampling on training folds only.

Performance was evaluated on a held-out time-based test set using accuracy, balanced accuracy, recall, precision, F1-score, ROC-AUC, and average precision. Given the public health context, recall was prioritized to minimize missed outbreak predictions.

### 3.5 Spatial and Ablation Analysis

Feature group importance was assessed using tree-based grouped importance scores. Ablation experiments evaluated the marginal predictive contribution of lagged climate features, vulnerability variables, and full temporal outbreak-history features relative to simpler feature blocks. State-level outbreak rates and burden profiles were used to characterize spatial concentration.

## 4. Results

### 4.1 Spatial and Temporal Outbreak Patterns

Cholera burden in the panel was highly concentrated both spatially and temporally. Among states, Katsina had the highest outbreak rate (0.605), followed by Zamfara (0.535), Bayelsa (0.535), Bauchi (0.512), and Kano (0.512). The heaviest total case burden was recorded in Borno (10,267 cases) and Bauchi (9,859 cases). Regionally, the North-West and North-East carried the most consistent burden, with outbreak rates of 0.412 and 0.407 respectively. The

North-East had the highest mean cases per outbreak month (241.9). Southern regions showed lower but non-negligible outbreak rates, with South-South at 0.213 and South-West at 0.182.

Temporally, outbreak activity peaked in July (outbreak rate = 0.411), August, and September. The highest total monthly case count was recorded in September (10,799 cases). Outbreak severity, measured as the share of high-severity months, was greatest in September (0.168), with elevated shares also in August and October. Year-level patterns showed sharp heterogeneity: 2021 had the largest total burden (20,668 cases and 106 outbreak months), while 2018 had the lowest (229 cases, outbreak rate = 0.108). The year 2025 had the highest observed outbreak rate (0.514) but reflects a partial year only.

### 4.2 Comparative Profiles: Outbreak Versus Non-Outbreak Months

The strongest standardized differences between outbreak and non-outbreak months were dominated by temporal outbreak-persistence variables. The largest contrasts were: lagged current cases log-transformed (standardized difference = 2.359), lag-1 outbreak flag (0.893), rolling 3-month outbreak sum (0.826), and prior state outbreak rate (0.599). WASH and poverty variables also differentiated outbreak months, with improved water access showing a negative contrast (−0.455), poor WASH index positive (0.406), poverty rate positive (0.392), and improved sanitation access negative (−0.336). These findings confirm that outbreak months are distinguished by both endemic persistence and structural vulnerability.

High-risk states differed from low-risk states by an outbreak rate gap of 0.344, a poverty rate difference of +36.6 percentage points, an improved water access difference of −16.2 percentage points, and an improved sanitation access difference of −10.3 percentage points. Notably, population density was lower in high-risk states (difference = −536.5), indicating that structural deprivation rather than urban density drives recurrent outbreak risk.

### 4.3 Feature Importance and Ablation Evidence

Grouped tree-based importance analysis identified outbreak-history as the dominant predictive feature group (importance share = 0.465), followed by structural vulnerability (0.181), geography (0.111), short-term climate (0.105), seasonality (0.054), interaction terms (0.054), and surface water (0.030).

Ablation experiments confirmed that lagged climate features added only modest incremental value: for Logistic Regression, lagged climate versus same-month climate improved ROC-AUC by 0.006 and average precision by 0.009; for Extra Trees, the gain was −0.007 in ROC-AUC and +0.008 in average precision. The strongest predictive lift came from adding full temporal outbreak-history features: for Logistic Regression, this improved ROC-AUC by 0.080 and average precision by 0.098; for Extra Trees, ROC-AUC improved by 0.074 and average precision by 0.062. Vulnerability variables showed mixed results when added in isolation but contributed positively in combination with richer temporal structure.

### 4.4 Model Comparison

Table 1 presents held-out test-set performance across the five primary model families.

**Table 1: Held-out model performance on the time-based test set**

| Model | Balanced Acc. | Recall | Precision | F1 | ROC-AUC |
|---|---|---|---|---|---|
| CatBoost | 0.530 | 0.991 | 0.387 | 0.556 | 0.668 |
| XGBoost | 0.519 | 1.000 | 0.381 | 0.551 | 0.654 |

| | | | | | |
|---|---|---|---|---|---|
| LightGBM | 0.514 | 0.991 | 0.378 | 0.548 | 0.646 |
| Logistic Regression | 0.515 | 0.955 | 0.379 | 0.543 | 0.655 |
| Random Forest | 0.500 | 1.000 | 0.372 | 0.542 | 0.677 |

CatBoost achieved the best balanced F1 performance (0.556), while Random Forest achieved the highest ROC-AUC (0.677). Recall remained very high across all models, supporting their usefulness for high-sensitivity early warning. Precision was more modest, reflecting the difficulty of predicting outbreak events in a heterogeneous panel with a 26.8% base rate.

### 4.5 Alert Outputs

Translating the latest CatBoost predicted probabilities into three alert categories yielded: CRITICAL (30 states), ELEVATED (6 states), and LOW (1 state). States classified as CRITICAL included Bayelsa, Delta, Rivers, Akwa Ibom, Abia, Kogi, Nasarawa, Zamfara, Benue, Borno, and Kano, among others. The high proportion of CRITICAL classifications reflects the model's recall-oriented threshold calibration and should be interpreted as a preparedness signal rather than a definitive forecast of imminent outbreak.

## 5. Discussion

This study developed and evaluated a probabilistic machine learning pipeline for cholera outbreak early warning in Nigeria, integrating climate, WASH, poverty, surface-water, and temporal surveillance features across a 2018–2025 state-month panel. Four findings merit discussion.

First, outbreak-history features substantially outperformed climate variables as predictors of one-month-ahead outbreak risk. The grouped importance analysis showed that temporal persistence features accounted for nearly half of all predictive signal, while climate contributed roughly one-tenth. This finding is consistent with the endemic nature of cholera in Nigeria's high-burden states, where recurrent outbreaks reflect deeply embedded structural conditions rather than transient climatic shocks [Mukandavire et al., 2011; Finger et al., 2014]. It implies that early-warning systems anchored solely in climate data will systematically underperform relative to systems that integrate surveillance history.

Second, structural vulnerability indicators — poverty, WASH access, and their interactions with climate — contributed meaningful secondary signal but did not yield strong isolated predictive gains in simple ablation blocks. Their strongest value appeared in combination with temporal outbreak-history features. This pattern is consistent with broader evidence that vulnerability shapes the background risk environment within which climatic triggers operate [Ali et al., 2015; Moore et al., 2017], rather than independently determining outbreak timing.

Third, predictive performance was moderate rather than high. The mean ROC-AUC of 0.659 and mean F1 of 0.547 across model families are honest reflections of the difficulty of predicting binary outbreak events one month ahead in a heterogeneous national panel. The very high recall (mean = 0.991) indicates that the system will rarely miss a true outbreak, which is the operationally critical property for preparedness planning. However, the modest precision means that false alarms are frequent, and the alert system should be communicated accordingly to decision-makers.

Fourth, the spatial concentration of burden in the North-West and North-East, and the strong seasonal peak in July through September, provide actionable targeting information that does not depend on model precision. Even without probabilistic scoring, these descriptive findings support pre-positioning of response resources in high-burden states before the peak season.

Several limitations apply. The study relies on cross-sectional monthly surveillance aggregates, which may undercount true outbreak burden due to variable reporting completeness across states and years. Static or slowly varying structural variables such as poverty and WASH access were repeated across monthly rows and do not capture intra-annual variation. GPS displacement in some underlying data sources introduces spatial imprecision. The framework is not yet deployment-ready and should be treated as a reproducible research prototype rather than an operational tool.

## 6. Conclusion

This study demonstrates that a reproducible probabilistic machine learning pipeline can generate meaningful early-warning estimates for cholera outbreak risk across Nigerian states. The system achieves high recall at the cost of modest precision, making it best suited to preparedness-oriented early warning rather than high-specificity operational triage. The dominant finding — that temporal outbreak-history features contribute more predictive signal than climate variables alone — has direct implications for how future cholera early-warning systems should be designed and validated. Structural vulnerability indicators provide important contextual signal but are most valuable in combination with surveillance history. Future work should explore the integration of real-time WASH and health system capacity data, finer spatial resolution, and prospective validation to move this framework toward operational readiness.

## References


Ali, M.; Nelson, A. R.; Lopez, A. L.; and Sack, D. A. 2015. Updated global burden of cholera in endemic countries. *PLOS Neglected Tropical Diseases* 9(6): e0003832. https://doi.org/10.1371/journal.pntd.0003832

Bertuzzo, E.; Finger, F.; Mari, L.; Gatto, M.; and Rinaldo, A. 2010. On the probability of extinction of the Haiti cholera epidemic. *Stochastic Environmental Research and Risk Assessment* 24: 1283–1296. https://doi.org/10.1007/s00477-010-0450-9

Bolu, O.; Ohuabunwo, C.; Ndodo, N.; Aworabhi-Oki, N.; Nguku, P.; Waziri, N. E.; Shuaib, F.; and Mwansa, J. 2019. Cholera epidemiology in Nigeria, 2012–2017. *Journal of Infectious Diseases* 218(Suppl. 3): S239–S246. https://doi.org/10.1093/infdis/jiz359

Brownstein, J. S.; Freifeld, C. C.; and Madoff, L. C. 2008. Digital disease detection — harnessing the web for public health surveillance. *New England Journal of Medicine* 360(21): 2153–2157. https://doi.org/10.1056/NEJMp0900702

Colwell, R. R.; Huq, A.; Islam, M. S.; Aziz, K. M. A.; Yunus, M.; Khan, N. H.; Mahmud, A.; Sack, R. B.; Nair, G. B.; Chakraborty, J.; Sack, D. A.; and Russek-Cohen, E. 2020. Reduction of cholera in Bangladeshi villages by simple filtration. *Proceedings of the National Academy of Sciences* 100(3): 1051–1055. https://doi.org/10.1073/pnas.0237386100

Finger, F.; Bertuzzo, E.; Luquero, F. J.; Naibei, N.; Touré, B.; Allan, M.; Porten, K.; Lessler, J.; Rinaldo, A.; and Azman, A. S. 2014. The potential impact of case-area targeted interventions in response to cholera outbreaks. *PLOS Medicine* 15(2): e1002509. https://doi.org/10.1371/journal.pmed.1002509

Jutla, A.; Whitcombe, E.; Hasan, N.; Haley, B.; Akanda, A.; Huq, A.; Alam, M.; Sack, R. B.; and Colwell, R. 2013. Environmental factors influencing epidemic cholera. *American Journal of Tropical Medicine and Hygiene* 89(3): 597–607. https://doi.org/10.4269/ajtmh.12-0721

Moore, S. M.; Azman, A. S.; Zaitchik, B. F.; Mintz, E. D.; Brunkard, J.; Legros, D.; Hill, A.; McKay, H.; Luquero, F. J.; Olson, D.; and Lessler, J. 2017. El Niño and the shifting geography of cholera in Africa. *Proceedings of the National Academy of Sciences* 114(17): 4436–4441. https://doi.org/10.1073/pnas.1617218114

Mukandavire, Z.; Liao, S.; Wang, J.; Gaff, H.; Smith, D. L.; and Morris, J. G. 2011. Estimating the reproductive numbers for the 2008–2009 cholera outbreaks in Zimbabwe. *Proceedings of the National Academy of Sciences* 108(21): 8767–8772. https://doi.org/10.1073/pnas.1019712108

Nigeria Centre for Disease Control (NCDC). 2022. *Cholera Situation Report Nigeria.* Abuja: NCDC. https://ncdc.gov.ng/diseases/sitreps

Rinaldo, A.; Bertuzzo, E.; Mari, L.; Righetto, L.; Blokesch, M.; Gatto, M.; Casagrandi, R.; Murray, M.; Vesenbeckh, S. M.; and Rodriguez-Iturbe, I. 2012. Reassessment of the 2010–2011 Haiti cholera outbreak and rainfall-driven multiseason projections. *Proceedings of the National Academy of Sciences* 109(17): 6602–6607. https://doi.org/10.1073/pnas.1203333109

World Health Organization (WHO). 2023. *Global Cholera and Acute Watery Diarrhoea Dashboard.* Geneva: WHO. https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON437