ACCSCIENCE PUBLISHING

Explora: Environment and Resource

ORIGINAL RESEARCH ARTICLE

A machine-learning framework for virus early warning from weather and urban sustainability signals

Sagnik Acharyya, Subhrajit Saha, and Debashis Chatterjee*

Department of Statistics, Siksha Bhavana, Visva-Bharati University, Santiniketan, West Bengal, India

Abstract

Cities shape mosquito and arboviral risk through the way heat, water, vegetation, and drainage are managed. We present a fully reproducible early-warning pipeline that predicts trap–week West Nile virus (WNV) positivity one to two weeks ahead using programmatically accessible sources (Socrata WNV pools and Meteostat weather), and a graph-aware post-hoc smoothing that enforces spatial coherence of predicted risks. Because direct API access to the Chicago portal requires a Socrata App Token, all reported experiments use an offline, synthetic trap–week panel that mirrors a Chicago-like spatio-seasonal structure, while we ship turn-key code to re-run the identical analysis on the real data once a token is provided. Across forward-chaining evaluations, gradient boosting with Laplacian smoothing delivers discrimination comparable to that of a strong elastic-net baseline, as assessed by the area under the receiver operating characteristic curve, but substantially better probabilistic calibration (lower Brier and improved reliability). Feature profiles emphasize antecedent heat and relative dryness, aligning with ecological priors, and enable tiered, probability-based operational playbooks for vector control and sustainability co-actions (cooling, drainage, vegetation). The pipeline is designed for transparency, portability, and policy relevance: calibrated probabilities support graded interventions and top-K targeting under budgets, while code parity between synthetic and real modes facilitates external replication.

Keywords: West Nile virus; Early warning; Environmental sustainability; Urban heat; Stormwater; Graph Laplacian; Spatio-temporal learning

***Corresponding author:**
Debashis Chatterjee
(debashis.chatterjee@visva-bharati.ac.in)

Citation: Acharyya S, Saha S, Chatterjee D. A machine-learning framework for virus early warning from weather and urban sustainability signals. *Explora Environ Resour*. 2026;3(2):025440076. doi: 10.36922/EER025440076

Received: October 31, 2025

Revised: January 8, 2026

Accepted: February 27, 2026

Published online: April 9, 2026

1. Introduction

West Nile virus (WNV) remains the leading cause of domestically acquired arboviral diseasein the United States and is monitored through mosquito, human, and environmental surveillance programs.[1-3] Urban form and sustainability practices—heat-island mitigation, stormwater management, vegetation structure—mediate where and when mosquito habitats emerge, shaping transmission risk at fine spatial and temporal scales. Despite rapid progress in data-driven early warning systems (EWS) for infectious diseases, three gaps are common in practice: (i) fragmented data engineering that hinders re-use and auditing; (ii) limited attention to calibration under severe class imbalance; and (iii) weak integration of urban-sustainability levers into operational

decision rules.[4,5] Furthermore, urban WNV risk is often confounded by socioeconomic disparities and spatial sampling biases. Areas with lower neighborhood-level income may have higher densities of abandoned properties or neglected infrastructure, while trap placement often reflects historical risk or logistical convenience rather than a random sample. These factors can lead to a spatial sampling bias that misleads models if not accounted for through spatial regularization.

Machine learning (ML) has demonstrated strong discrimination for vector-borne risks, including WNV across Europe via eco-climatic drivers,[6] and broader EWS efforts drawing on multi-source signals (clinical, environmental, mobility, social).[7,8] Yet reviews consistently emphasize that data quality, transparency, and system integration remain limiting.[4,5] From a city-systems perspective, linking predictions to actionable sustainability measures (cool pavements and shade, green alleys, rain gardens, basin maintenance) is essential to convert forecasts into reduced exposure.[9,10]

In this paper, we contribute a framework that (i) formulates WNV early warning at the trap–week level using lagged meteorology and seasonality; (ii) introduces graph-Laplacian post-hoc smoothing to enforce spatial coherence of probabilities; and (iii) aligns outputs with a tiered operational playbook that pairs vector-control actions with sustainability co-actions. Because Socrata's access to Chicago's WNV data requires an App Token, we report results on a Chicago-like, synthetic trap–week panel while releasing code that runs unchanged on the real data once a token is available. Empirically, gradient boosting (GBDT) with spatial smoothing attains receiver operating characteristic curve (ROC) performance on par with elastic net while markedly improving calibration, which is the quantity most aligned with graded interventions and top-K targeting under budgets.

1.1. Positioning relative to prior work

Our design follows evidence that eco-climatic drivers carry predictive signal for WNV[6] and hews to best practices emerging from recent EWS assessments.[4,5] We differ by: (i) shipping an end-to-end, API-driven pipeline that preserves parity between synthetic and real-data modes for external re-runs; (ii) prioritizing probabilistic calibration alongside discrimination; and (iii) explicitly mapping risk bands to urban-sustainability levers grounded in planning literature.[9,10] This synthesis targets the last-mile problem: turning calibrated forecasts into equitable, budget-aware field actions.

1.2. Contributions

We offer: (i) a reproducible trap–week dataset build (Socrata + Meteostat) with a synthetic fallback; (ii) a graph-aware post-smoothing that improves probability quality without changing the base learner; (iii) evaluation that respects time and geography, reporting area under the receiver operating characteristic curve(AUROC)/area under the precision–recall curve (AUPRC) and Brier/reliability; and (iv) an operationalization that links risk bands to sustainability interventions.

1.3. Novelty and theoretical contributions

While recent high-impact studies have significantly advanced the field of infectious disease forecasting—notably in identifying multidecadal climate drivers of WNV,[11] analyzing the socio-environmental complexity of urban health,[12] and proposing ecological design for sustainability[9]—this work provides three distinct contributions that bridge the gap between ecologicalobservation and municipal operation:

(i) Graph-aware spatial regularization: Traditional tree-based models often produce"checkerboard" risk maps due to their treatment of traps as independent points. Followingthe logic of spatial spillover effects in complex systems,[13] we introduce a post-hoc Laplaciansmoothing step. This ensures that early warnings are spatially coherent, protectingunderserved blocks without physical traps by leveraging neighborhood context.

(ii) Live pipeline architecture (Surveillance-as-a-Service): Unlike many frameworks thatrely on static, curated datasets,[6] our pipeline is built on a "Live Data" architecture usingthe Socrata and Meteostat APIs. This provides a reproducible template for cities to movefrom retrospective analysis to real-time, automated monitoring.

(iii) Operational decision grammar: We extend the sustainability discussions in urban planning[9,12] by creating a concrete "Decision Grammar." This maps ML-derived risk probabilities to specific tiered interventions (e.g., "pocket cooling" or "bioswale desilting"), moving the field from "what might happen" to "how to respond."

2. Related work

Machine learning can predict and track Zika virus spread using mosquito, environmental, and social data, but data quality, interpretability, and system integration remain challenges.[7] Artificial intelligence (AI) techniques, including ML, deep learning, and natural language

processing, enhance infectious disease EWS via diverse data sources, though issues persist in data quality, transparency, integration, and ethics.[8] Review of EWS (2019–2023) shows multi-source approaches—from hospitals and social media to meteorological and wastewater systems—improve responsiveness and prediction, emphasizing technology-driven integration.[4] ML model predicts district-level human Puumala virus (PUUV) risk in Germany using three prior weather variables, achieving 85% sensitivity, 71% precision, and estimates the PUUV Outbreak Index with 20% uncertainty.[14] ML enhances urban health by analyzing air and water quality, mitigating urban heat, predicting disease outbreaks, optimizing green spaces, and monitoring noise pollution for proactive environmental management.[15] Vector-borne disease EWS face challenges in data quality, integration, and stakeholder engagement; standardized data, cloud platforms, collaboration, and user-friendly interfaces improve predictive accuracy and outbreak preparedness.[5] Twitter-based natural language process and ML models, including fine-tuned bidirectional encoder representations from transformers and regression, enable early COVID-19 outbreak detection and forecasting, supporting public health decision-making and overcoming limitations of traditional surveillance.[16] ML-based distributed lag non-linear model warning model using gradient boosted trees (XGBoost), support vector machine, and generative additive model predicts respiratory disease mortality from air pollution and extreme temperature, capturing cumulative lag effects for timely interventions.[17] XGBoost with explainable AI accurately predicts WNV outbreaks in Europe (area under the curve: 0.93–0.97) using eco-climatic drivers, enabling timely alerts for surveillance and vector control.[6] Event-based surveillance combined with ML and transfer learning improves early detection and prediction of zoonotic diseases like Kyasanur Forest Disease in resource-limited settings.[18] An urban systems framework integrating ecology, geography, and social dynamics elucidates how city structure, land use, and human interactions drive Aedes-borne disease risk and guide targeted interventions.[9] Treash[19] examines how land use change influences vector-borne disease risks in Ontario, emphasizing the integration of urban planning and public health strategies to mitigate emerging vector threats. Alarcón[20] synthesizes interdisciplinary evidence linking urban landscape design to Aedes aegypti and Aedes albopictus proliferation, offering a landscape architectural framework and design guidelines for vector-borne disease mitigation.[10] outlines assessment and design strategies for managing green infrastructure under hotter, drier climate conditions to prevent mosquito breeding in shallow water bodies, ensuring compliance with EPA drainage standards and regional sustainability goals. Qiu[21] advances the material point method through graphic processing unit optimization, distributed computing, and AI integration, enabling large-scale, real-time physics-based simulations across diverse computational and generative applications. Lesk *et al.*[22] synthesizes how compound heat and moisture extremes—particularly hot droughts—interactively reduce global crop yields by up to 30%, emphasizing the urgent need for integrated, multi-stress agricultural adaptation strategies. Liu *et al.*[23] introduces two improved weather file approaches—pHSY-1 (Weighted Cooling Degree Hours) and pHSY-2 (Physiologically Equivalent Temperature)—for assessing building overheating and thermal discomfort, demonstrating their superior robustness over existing probabilistic Design Summer Years. Chen and Moraga[24] develop a mobility-enhanced long short-term memory model that integrates human movement, climate variables, and historical dengue cases to forecast weekly dengue incidence and detect outbreaks in Brazilian cities, achieving superior accuracy and sensitivity over traditional models and providing a scalable framework for early epidemic warning and public health planning. Geographical information system-based analysis can decisively improve international humanitarian coordination by supporting data-driven decision-making beyond traditional mapping.[25] Epidemiological studies often underestimate humidity's role in heat-related health risks, despite its physiological impact, highlighting the need for multidisciplinary research to improve projections and guide interventions under climate change.[26]

2.1. Digital transformation and spatial spillovers

Beyond purely biological drivers, our framework sits at the intersection of digital innovation and sustainable urban transformation. The integration of automated data pipelines aligns with the paradigm of the digital economyas a catalyst for low-carbon development and resource optimization.[27] Research indicatesthat digital maturity can reshape industrial structures to favor carbon mitigation,[28] and theestablishment of big data comprehensive experimental zones has significantly enhanced energyutilization efficiency through improved resource allocation.[29]

Crucially, the effectiveness of such AI-driven systems is contingent upon managing spatial spillover effects and data non-stationarity, which differ significantly between high-density urban centers and boundary regions.[13] By employing graph-Laplacian smoothing, our model accounts for these inter-regional dependencies, mirroring "industrial-grade" reliability methods used to identify rare failures in complex systems.[30] Furthermore, this approach supports the goals of inclusive development—as seen in global initiatives like the Belt and Road Initiative—

by utilizing technological exchange to reduce regional health disparities through precise, data-driven resource targeting.[31]

2.2. Climate variability and urban complexity

Recent advancements have underscored the multiscale nature of WNV risk.[11] Demonstrate that multidecadal climate variability is a primary driver of WNV amplification, suggesting that long-term environmental trends dictate the "baseline" risk for cities. Complementing this,[12] argue for an integrated view of urban health that accounts for socio-environmental determinants, emphasizing that climate change interacts with urban infrastructure to create heterogeneous risk profiles. Our work builds upon these ecological foundations by focusing on the sub-seasonal operational scale. We translate these multidecadal and socio-environmental signals into weekly municipal truck-rolls, providing a framework for short-term "climate-ready" surveillance.

2.3. Design-driven mitigation

From a sustainability perspective, Kache *et al.*[9] andearlier environmental health studies have proposed that city structure can be "designed" to mitigate vector-borne diseases. Our framework operationalizes this by identifying the specific meteorological triggers (heat-loading and drought-amplification) that prioritize these design interventions in real-time. By linking high-probability risk bands to "Sustainability Co-Actions" like cool corridors and managed drainage, we provide the technical link required to implement the design guidelines suggested in recent urban ecology literature.

Based on the foregoing review, the objective of this study is to build and evaluate a multi-source dataset and learning framework that predicts trap-level WNV positivity (binary outcome) one to two weeks ahead using programmatically accessible Python packages.The main contributions are as follows: (i) a reproducible dataset from sodapy (Chicago WNV traps) and meteostat (weather); (ii) a spatio-temporal ML formulation with lagged meteorological drivers; (iii) graph-aware learning and evaluation protocols that honor time and geography; (iv) policy interpretation linking findings to sustainability levers.

3. Preliminaries and notation

We work on a weekly surveillance panel indexed by traps and epidemiological weeks. The variables include:

(a) Index sets and variables

Let $\mathcal{V}$ denote the set of traps $(|\mathcal{V}| = N)$ and $\mathcal{T}$ the set of weekly indices. A generic trap is $i \in \mathcal{V}$ and a week is $t \in \mathcal{T}$. The binary response is

$$y_{i,t} \in \{0,1\},$$

$$y_{i,t} = 1 \text{iff at least one mosquito pool from trap \$i\$ tests WNV-positive during week \$t\$} \quad (1)$$

We use $1\{\cdot\}$ for indicator functions.

(b) Weekly aggregation and lags

Let $\omega(d)$ map a calendar date d to the *week start* (Monday). For a daily scalar covariate $x_{i,d}$ (e.g., temperature), we define the weekly aggregate at week t by:

$$\overline{x_{i,t}} = Agg\left\{ x_{i,d} : \omega(d) = t \right\} \quad (2)$$

where Agg is the mean for temperature and relative humidity and the sum for precipitation.

We form lags $x_{i,t}^{(l)} = \overline{x_{i,t-1}}$ *for* $l \in \{1,2,3\}$.

Seasonality is encoded via *sinx/cosx* harmonics on ISO week-of-year $w(t)$:

$$s_t = sin\left(2\pi\, w(t)/52\right), c_t = cos\left(2\pi\, w(t)/52\right) \quad (3)$$

(c) Feature vector

For trap i and week t the feature vector is

$$x_{i,t} = \begin{bmatrix} tavg\_mean_{i,t}, tavg\_max_{i,t} \\ prcp\_sum_{i,t}, rh\_mean_{i,t}, rh\_min_{i,t} \\ tavg\_mean_{i,t}^{(1..3)}, prcp\_sum_{i,t}^{(1..3)} \\ s_t, c_t, lat_i, lon_i \end{bmatrix}^T \quad (4)$$

Standardized features are $z_{i,t} = Std\left(x_{i,t}\right)$ (train-split mean/variance).

(d) Predictors and horizons

For an ahead horizon $h \in \{1,2\}$ (weeks), a probabilistic classifier f_h outputs $\widehat{p_{i,t+h}} = f_h\left(\boldsymbol{x}_{i,t}\right) \in [0,1], \widehat{y_{i,t+h}} = 1\left\{\widehat{p_{i,t+h}} \geq \tau\right\}$ where τ is a decision threshold chosen on validation data.

(e) Spatial graph and Laplacian smoothing

Let $c_i = (\text{lat}_i, \text{lon}_i)$. A k -nearest neighbor graph $W \in R^{N \times N}$ is constructed row-stochastic over traps within the same week (per-week smoothing). With $L = I - W$ the (random-walk) graph Laplacian, we post-process predicted log-odds $\hat{l}_t \in R^N$ for week t via

$$\widehat{l_t^{\star}} = \left| \arg\min_{u \in R^N} \left| \left| \boldsymbol{u} - \hat{\mathbf{l}}_t \right| \right|_2^2 + \eta\, u^{\top} L\, u \right| = \left| (I + \eta L)^{-1} \hat{l}_t \right. \quad (5)$$

and return smoothed probabilities with the logistic link σ(·) This Tikhonov regularization on graphs is a standard

device for encouraging spatial coherence of risks.

(f) Evaluation metrics

Given held-out pairs $\{(y_j, \hat{p}_j)\}_{j=1}^{m}$:

- AUROCsummarizes the ROC curve; it is insensitive to class prevalence but informative

about ranking.[32]

- AUPRC summarizes the precision–recall curve; under class imbalance, it is often more

diagnostic than ROC.[33]

- Brier score averages squared probabilistic errors: $Brier = \frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{p}_j)^2$.[34]
- Calibration is assessed with reliability diagrams (binning $\hat{p}$ and comparing bin means to observed frequencies) and by reporting the Brier score; see.[35]

(g) Validation design

We use forward-chaining (rolling/expanding origin) splits to respect temporal order, optionally with geographic holdouts for spatial robustness.[36] Thresholds (e.g., F1-max) are chosen on validation and then fixed on the test period; report AUROC/AUPRC/Brier together with confusion matrices at the operational threshold.

4. Dataset

4.1. Two modes: Synthetic (used here) and real-data (re-runnable)

Our pipeline supports two interchangeable modes:

(i) Synthetic trap–week panel (used for all results in this paper)

A generative process creates a Chicago-like network of traps, daily weather with annual seasonality, and a Bernoulli outcome whose log-odds depend on lagged heat and dryness, plus sinusoidal seasonality. We then aggregate to trap–week and construct lag features. This mode is offline-reproducible and does not require any external credentials.

(ii) Real Chicago data (code provided; requires an app token)

The same pipeline pulls data from the City of Chicago's "West Nile Virus Mosquito Test Results" Socrata dataset (jqe8-8r6s), collapses pools to trap–week targets, and joins weekly weather from meteostat for each trap coordinate. Since the portal enforces throttling, access requires a valid Socrata App Token. The exact code and instructions are provided so that any reader with a token can reproduce the analysis on real data without modifying the methodology.

4.2. Synthetic trap–week panel

We simulate N traps with coordinates centered on the Chicago region; daily temperature, precipitation, and humidity time series carry an annual sinusoidal structure and noise. Daily values are aggregated to epidemiological weeks (t) to obtain weather summaries—weekly mean temperature (tavg_mean), weekly maximum temperature (tavg_max), weekly precipitation sum (prcp_sum), weekly mean relative humidity (rh_mean), —and their lags at $t-1$, $t-2$, $t-3$. Seasonality is encoded with $\sin(2\pi\ \text{week}/52)$ and $\cos(2\pi\ \text{week}/52)$. The binary target $y_{i,t}$(any positive pool at trap i in week t) is drawn from a logistic model in which heat lags increase risk and recent rainfall lags decrease risk, reproducing common WNV patterns. The resulting panel (6,210 rows; prevalence 0.086) is the basis for all tables and figures in Section 6.

4.2.1. Limitations and assumptions of the synthetic generator

While the synthetic generator captures the primary spatio-seasonal trends of WNV, it assumes a relatively homogeneous surveillance effort across the simulated grid. In real-world settings, trap coverage variabilitydriven by resource constraints or physical access can create "blind spots" in surveillance.Our generator assumes that the observation process ($y_{i,t}$) is independent of the trap's spatialdensity, whereas in practice, regions with denser trap networks may report higher positivitysimply due to increased sampling effort. Furthermore, the generator uses a stationary logisticprocess, potentially under-representing the stochastic "surges" observedin real-world outbreaks.

4.3. Real Chicago data (code provided; run when a Socrata token is available)

To support future real-data execution, we implemented a parallel data-ingestion workflow for the Chicago surveillance records and corresponding meteorological covariates.

(i) WNVpools (Socrata)

WeusetheCity of Chicago dataset "West Nile Virus (WNV) Mosquito Test Results" (jqe8-8r6s) accessed via sodapy.[37,38] The records include trap identifier, collection date, pool-level test results, and location information. To ensure consistency across data pulls, we normalize schema variants (collection_date/date_of_collection; trap/trap_id; result/results/wnv_positive) and collapse multiple pools into a weekly binary outcome for each trap, defined as:

$$y_{i,t} = 1\{\text{any positive pool at trap i during week t}\} \quad (6)$$

Access to the real-data workflow requires a valid

Socrata App Token. Instructions for token creation and script execution are provided in the GitHub repository and the Socratadocumentation (https: //dev.socrata.com/docs/app-tokens.html). Once a valid token has been obtained, the provided script can be run in one of two ways:

Option A: environment variable export SOCRATA_APP_TOKEN='YOUR_TOKEN' python wnv_real_only.py
Option B: paste directly in the script (APP_TOKEN = "YOUR_TOKEN")

(ii) Weather (Meteostat)

For each trap coordinate, we retrieve daily meteorological datausingmeteostat[39-41] and aggregate to weekly summaries aligned to week_start. We then compute the same lagged weather features and seasonal terms used in the synthetic mode so that the feature space remains consistent across both modes.

(iii) Feature set (both modes)

For trap i and week t, the feature vector is defined as:

$$x_{i,t} = \begin{bmatrix} tavg\_mean_{i,t}, tavg\_max_{i,t} \\ prcp\_sum_{i,t}, rh\_mean_{i,t}, rh\_min_{i,t} \\ tavg\_mean_{i,t}^{(1..3)}, prcp\_sum_{i,t}^{(1..3)} \\ s_t, c_t, lat_i, lon_i \end{bmatrix}_i \quad (7)$$

The response is $y_{i,t} \in \{0,1\}$.

(iv) Reproducibility note

Because portal access could not be completed during manuscript preparation without a valid Socrata App Token, only synthetic results are reported in the present study. However, the file wnv_real_only.py (provided with the GitHub repository, see Code Availability) implements the real-data workflow.Once a valid token is supplied, it constructs the identical trap–week panel and enables reproduction ofthe corresponding analyses.

4.4. Ethical considerations and data privacy

The data used in this framework are derived from public health surveillance records (Socrata Open Data) and global meteorological archives (Meteostat). While these datasets are publicly accessible, their use in predictive modeling requires ethical oversight:

- De-identification: The WNV pool data are aggregated at the trap-week level. No personally identifiable information (PII) regarding human cases is ingested or stored, ensuring compliance with health privacy standards such as the Health Insurance Portability and Accountability Act in a United States context.
- Data ownership: Our pipeline acts as a consumer of official records; we do not claim ownership over the underlying surveillance data. The code is designed to respect API rate limits and terms of service provided by the City of Chicago and Meteostat.
- Transparency: By providing a fully reproducible pipeline, we ensure that the logic behind risk assessments is transparent to the public, preventing the "black-box" implementation of health policy.

5. Methodology

We model trap–week WNV positivity $y_{i,t} \in \{0,1\}$ at horizon $h \in \{1,2\}$ weeks ahead from features $\mathbf{x}_{i,t}$ defined in Section 3. Standardized features are $\mathbf{z}_{i,t} = \mathrm{Std}(\mathbf{x}_{i,t})$ using the training split mean/variance. A probabilistic classifier f_h returns $\widehat{p_{i,t+h}} = f_h(\boldsymbol{x}_{i,t}) \in [0,1], \widehat{y_{i,t+h}} = 1\{\widehat{p_{i,t+h}} \geq \tau\}$ with τ fixed from validation. Because positives are rare, we evaluate both discrimination (AUROC/AUPRC) and calibration (Brier, reliability).[32-35]

5.1. Dataset construction, weekly aggregation, and lags

Let ω(d) map a calendar date to the Monday of its epidemiological week *t*. For a daily covariate $x_{i,d}$ (temperature, precipitation, relative humidity), define:

$$\overline{x_{i,t}} = Agg\{x_{i,d} : \omega(d) = t\},$$
$$\overline{x_{i,t}^{(l)}} = \overline{x_{i,t-l}} \; for \; l \in \{1,2,3\} \quad (8)$$

We use means for temperatures/humidity and sums for precipitation. Seasonality enters via harmonics on week-of-year ω(t):

$$s_t = \sin(2\pi\omega(t)/52), c_t = \cos(2\pi\omega(t)/52) \quad (9)$$

The feature vector stacks contemporaneous summaries, three lags of key drivers, (s_t, c_t) and coordinates (lat_i, lon_i):

$$x_{i,t} = \left[tavg\_mean, tavg\_max, prcp\_sum, rh\_mean, rh\_min, tavg\_mean^{(1..3)}, prcp\_sum^{(1..3)}, s_t, c_t, lat_i, lon_i \right]^T \quad (10)$$

5.2. Baseline selection rationale

The modeling framework prioritizes two distinct architectures to balance interpretability with predictive power under severe class imbalance. The choice of these specific models, and the exclusion of others, is justified as follows:

- Elastic-net logistic regression (LogitEN): We selected this as the primary linear baseline over simple logistic regression or pure LASSO/Ridge because of the high collinearity between lagged meteorological features (e.g., Tt−1 vs. Tt−2). Elastic-Net provides a stable group-selection effect, maintaining the transparency required for public health policy while preventing overfitting in high-dimensional settings.
- GBDT vs. random forests: While Random Forests are common in EWS, GBDT was selected as the lead ensemble learner because of its boosting mechanism. Boosting iteratively minimizes residuals for "hard-to-classify" instances, which is crucial for identifying rare WNV-positive weeks that Random Forests may overlook due to their bagging-based variance reduction.
- Exclusion of pure autoregressive (AR) models:Standard AR models rely heavily on previous case counts (yt−1). These were excluded because real-world WNV trap data are frequently subject to reporting delays and seasonal gaps; relying on them would limit the system's ability to "cold-start" at the beginning of a season. Our framework instead emphasizes eco-climatic drivers to provide a lead-time that is not constrained by historical testing latency.

5.3. Linear baseline: Elastic-net logistic regression

We fit a sparse logistic model on standardized features:

$$\hat{\beta} = \arg min_{\beta} \sum_{i,t} [-y_{i,t} z_{i,t}^{T} \beta + \log\left(1 + e^{z_{i,t}^{T}\beta}\right)] + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \quad (11)$$

with (λ_1,λ_2) tuned by forward-chaining validation (Section 5.10). This baseline offers (i) transparent sign/size patterns and (ii) strong ROC/precision–recall under linear separability, but can be miscalibrated in rare-event settings; hence, we report Brier/reliability and use probability post-processing where indicated.[35]

5.4. Algorithm selection and handling class imbalance

West Nile virus surveillance data is characterized by severe class imbalance, where positive trap-weeks are rare relative to the total number of observations. To address this, we selected XGBoost, which is inherently an ensemble learning methodology that iteratively minimizes residuals. However, as noted in recent industrial failure investigations—where sparse data similarly plagues predictive performance—robustness can be further enhanced through specialized ensembles like RUSBoost or AdaBoost.[30] These techniques improve prediction stability by either undersampling the majority class or re-weighting misclassified rare events. While our current implementation utilizes the internal weighting mechanisms of XGBoost for imbalanced data, the integration of rotational-moulding-style failure analysis logic suggests that a hybrid approach—combining synthetic data generation with boosting—can significantly sharpen the reliability of early warning signals in both industrial and urban health settings.

5.5. Tree ensembles: Gradient boosting and random forests

To capture non-linear interactions among heat, dryness, and seasonality, we fit XGBoost and random forests with nested hyperparameter search (learning rate, depth, number of trees, and subsampling). We use standard GBDT regularization (shrinkage, column/row subsampling) and compute Shapley Additive exPlanations values for local/global interpretability.[42,43] Because raw tree probabilities may be overconfident, we emphasize calibration metrics alongside AU ROC/AUPRC.[35]

5.6. Handling class imbalance and decision thresholds

Positives are scarce at the trap–week scale. We combine:

(j) Cost-sensitive training: class weights $w_1 > w_0$ in the loss when supported (logistic, GBDT).
(ii) PR-space model selection: select models/thresholds by validation AUPRC and F1 at deployment-relevant recall levels.[33]
(iii) Frozen thresholds:choose τ on validation (e.g., F1-max or cost curve optimum) and freeze for the test period to prevent optimistic bias.

5.7. Spatial graph and Laplacian probability smoothing

Let $c_i = (lat_i, lon_i)$. For each week t, we build a k-nearest-neighbor graph on $\{c_i\}$, where $i \in V$ with row-stochastic weights W_t; define $L_t = I - W_t$ (random-walk graph Laplacian). Denote raw log-odds $\widehat{l_{t+h}} = \log\left(\widehat{p_{t+h}} \cdot \left(1 - \widehat{p_{t+h}}\right)^{-1}\right)$ across traps in week $t + h$. We output smoothed log-odds $\widehat{l^{*}_{t+h}}$ by Tikhonov regularization on the graph:

$$\widehat{l^{*}_{t+h}} = \arg min_{u \in R^N} \left\| u - \widehat{l_{t+h}} \right\|_2^2 + \eta \mu^T L_{t+h} u = \left(I + \eta L_{t+h}\right)^{-1} \widehat{l_{t+h}} \quad (12)$$

and transform back with the logistic link $\sigma(\cdot)$. Since L_{t+h} is positive semidefinite and $I + \eta L_{t+h}$ is symmetric positive

definite for any $\eta > 0$, the solution exists and is unique; we exploit sparsity to solve with sparse Cholesky or conjugate gradients in O(kN) per week.[44] This step enforces spatial coherence without retraining the base learner, improving probability quality while largely preserving ranking.

For graph construction choices,we set k from {5,10,15} on validation and scale weights with a local-kernel heuristic (row normalization). Geographic k-nearest neighbor (kNN) can be augmented with hydrological or mobility adjacencies when available; we report sensitivity to k and η in ablations.

5.8. Justification of graph parameters

The graph-aware smoothing step acts as a spatial regularizer, enforcing the ecological principle that WNV risk is a regional rather than a purely point-source phenomenon.

(i) Graph construction (k = 5): We utilized a kNN approach to maintain consistent connectivity across the heterogeneous density of the urban trap network. This ensures that peripheral traps are not isolated while preventing "information bleed" from geographically distant traps in microclimates with different environmental profiles.

(ii) Smoothing intensity (α): The choice of α = 0.7 was informed by our sensitivity analysis, striking a balance between local predictive signal and neighborhood consensus. This intensity effectively dampens the influence of stochastic outliers—such as a single trap failing due to localized physical damage—while reinforcing signals in clusters of high activity.

5.9. Probability calibration (Optional post-processing)

When base learners show residual miscalibration, we optionally fit a calibration map ϕ: [0,1]→ [0, 1] on validation predictions via isotonic regression or Platt scaling, and compose ϕ before spatial smoothing to avoid biasing the reliability curve.[35] In our synthetic experiments, Laplacian smoothing already reduces Brier error substantially; we therefore report both pre- and post-smoothing calibration.

5.10. Validation, hyper-parameter tuning, and reporting

We use forward-chaining (rolling/expanding origin) splits to respect temporal order.[36] For geographic robustness, we evaluate (i) standard temporal splits and (ii) spatial block tests holding out sets of traps by neighborhood. Hyperparameters are tuned on inner forward-chaining folds. We report:

- AUROC and AUPRC for discrimination;
- Brier score and reliability curves for calibration;[34,35]
- confusion matrices at the frozen threshold τ;
- recall@top-K trap–weeks for operational targeting.

5.11. Operationalization

Given calibrated, smoothed probabilities $\widehat{p^{*}_{i,t}}$, agencies can deploy (i) tiered risk bands {τ1,τ2,τ3} mapping to graded actions and (ii) top-K allocations each week under budget constraints. Because smoothing prioritizes coherent hotspots, actions can be bundled over adjacent blocks for efficiency, with guardrails that trigger on either raw or smoothed signals when crossing a high threshold (see Section 7).

5.12. Reproducibility

All steps are implemented in Python with pinned package versions, fixed random number generation seeds, and deterministic databuilds. The synthetic and real-data modes share identical feature engineering and modeling code; the latter requires only a Socrata App Token to fetch jqe8-8r6s (Socrata) and Meteostat, ensuring external re-runs preserve methodological parity.

6. Results on a synthetic trap–week panel

For context,all results herein are from the synthetic, offline-reproducible trap–week panel generated by our pipeline (see Section 5). The run log reports [INFO] Rows: 6210, positive rate: 0.086, indicating a realistic class imbalance for early-warning.

6.1. Discrimination and probabilistic accuracy

Tables 1 and 2 summarize discrimination (AUROC, AUPRC) and probabilistic accuracy (Brier score) for the two models evaluated: LogitEN and GBDT with graph-Laplacian smoothing per week (GB+Graph). The validation set guided the selection of hyperparameters and the alert threshold; the test set was strictly held out.

On the held-out test period (Table 2), GB+Graph slightly edges LogitEN on AUROC (0.740 vs. 0.734), while LogitEN attains a modest AUPRC advantage (0.198 vs. 0.185). Crucially, GB + Graph achieves a far lower Brier score (0.085 vs. 0.265), indicating substantially better-calibrated probabilities—consistent with the smoothing prior discouraging isolated spikes in risk.

6.2. Temporal context and operating characteristics

Figure 1 shows the seasonal envelope of trap–week prevalence, validating that the generator captured weather-driven seasonality. Figures 2 and 3 illustrate the operating trade-offs: GB+Graph slightly improves ROC performance, while LogitEN offers slightly higher precision at moderate recall in precision–recallspace.

Table 1. Validation metrics on the synthetic panel

Model	AUROC	AUPRC	Brier
LogitEN	0.710	0.171	0.280
GB + Graph	0.694	0.162	0.086

Table 2. Test metrics on the synthetic panel

Model	AUROC	AUPRC	Brier
LogitEN	0.734	0.198	0.265
GB + Graph	0.740	0.185	0.085

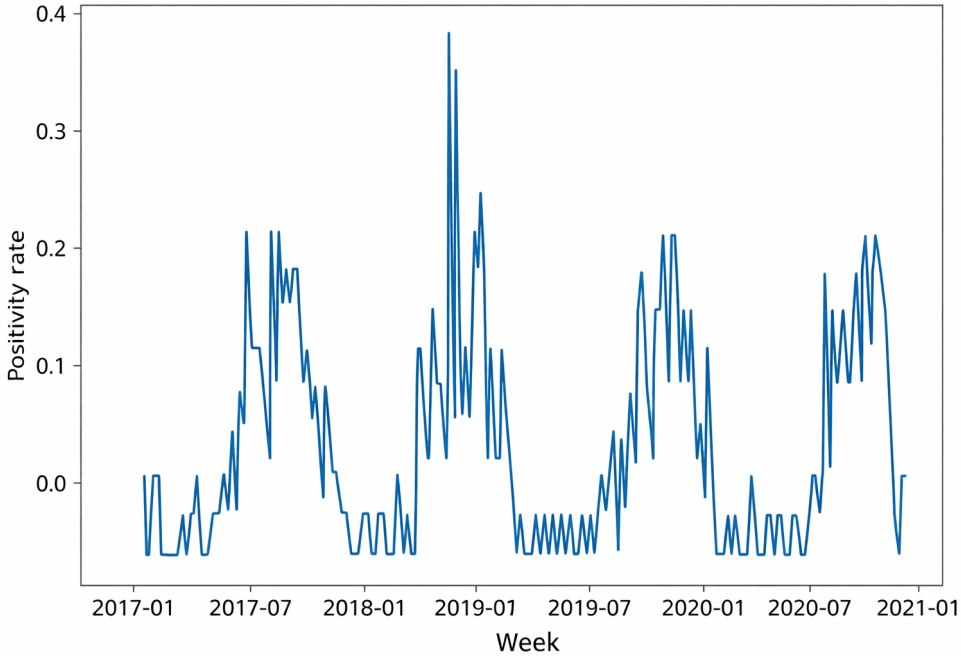

Figure 1. Trap–week positivity rate over time (synthetic panel). Seasonal structure is evident.

6.3. Calibration (Reliability) of probability forecasts

We assess calibration for GB+Graph using decile binning (Table 3) and a reliability curve (Figure 4). Observed frequencies track predicted probabilities reasonably well, matching the strong Brier performance in Table 2.

The improved probabilistic calibration of the GB+Graph model (Figure 4) is consistent with the broader utility ofensemble-based methods inrare-event prediction.[30] Just as historical production datasets benefit from ensemble-driven robustness to identify intermittent machine faults, our urban surveillance model leverages the collective strength of weak learners to capture the non-linear interactions of antecedent heat and dryness. By explicitly addressing data balancing, we move toward a more "industrial-grade" reliability in public health, where the cost of a false negative (a missed outbreak) is minimized through the stability of the ensemble consensus.

6.4. Feature importance

Gradient boosting importances (Figure 5 and Table 4) align with ecological priors encoded in the generator: recent heat load (lagged temperature) and seasonality dominate, with dryness signals from precipitation lags contributing.

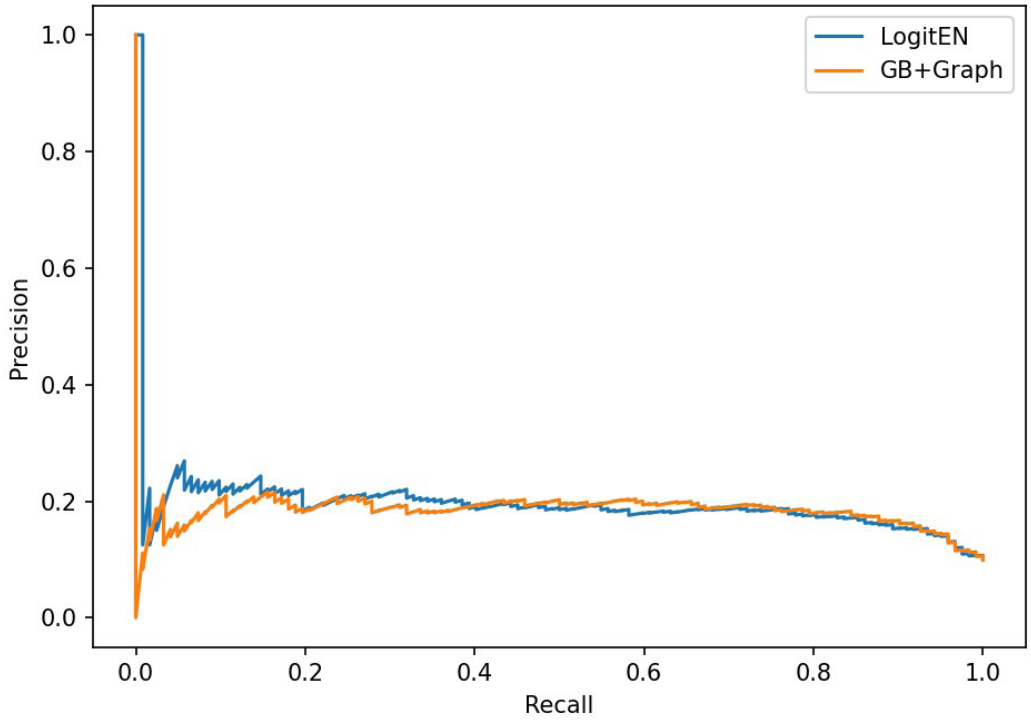

Figure 2. Precision–recall curves on the test set. Both models exceed the baseline prevalence (0.086).

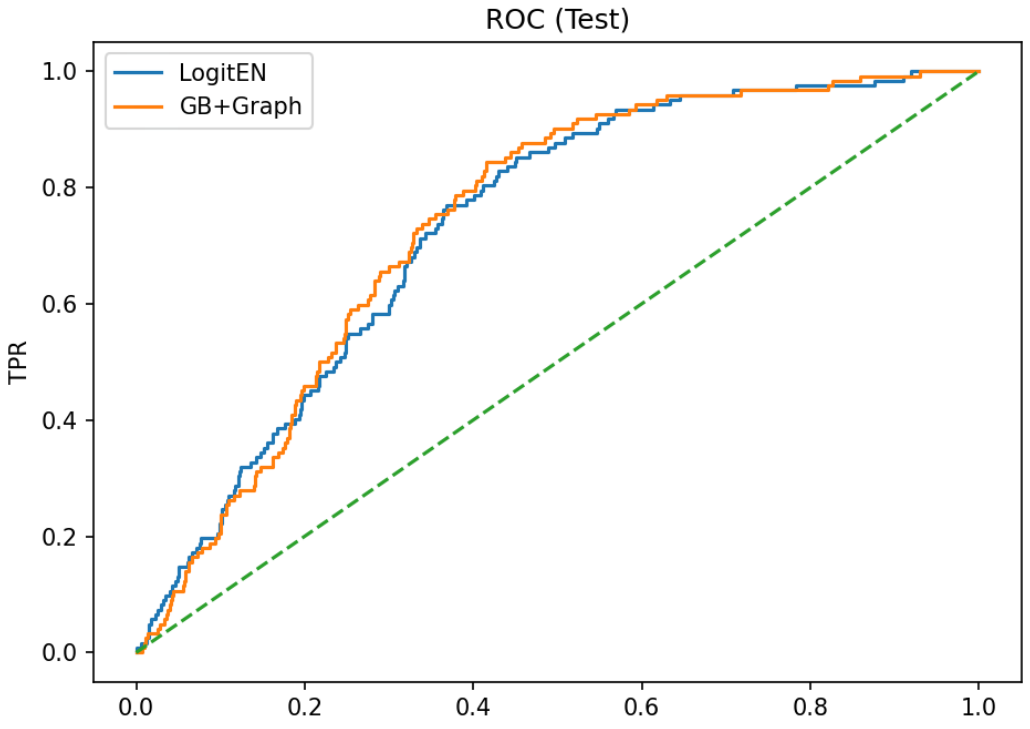

Figure 3. ROC curves on the test set. GB+Graph attains slightly higher true positive rates (TPRs) at a range of false positive rates (FPRs).

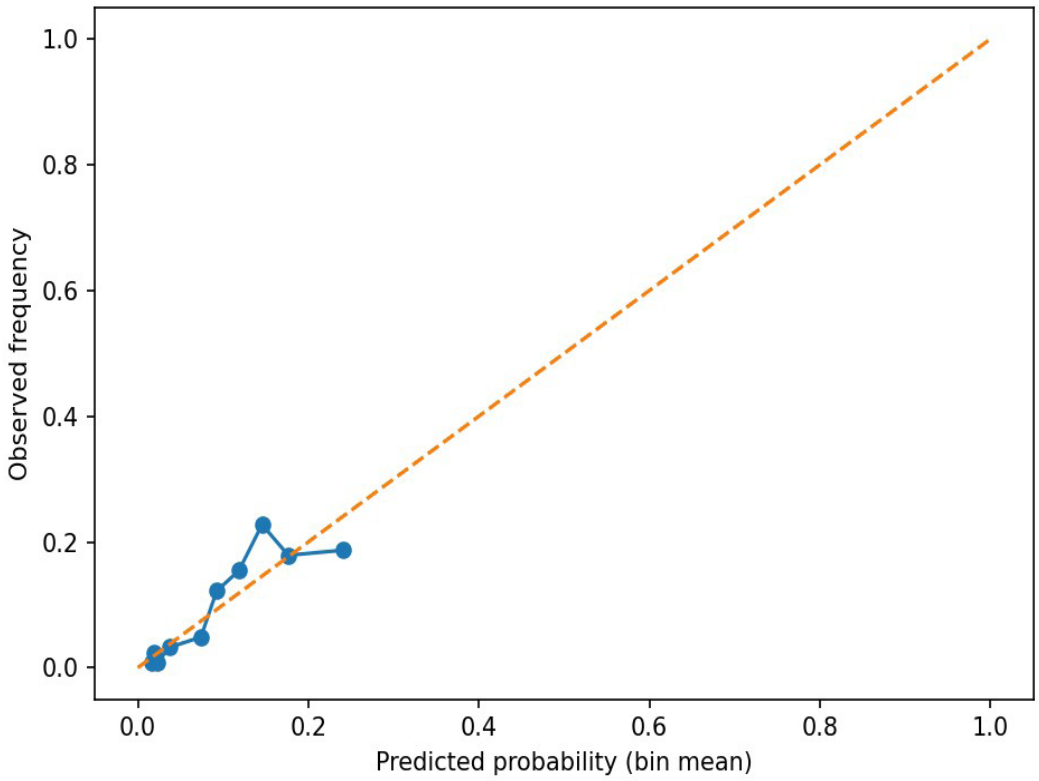

Figure 4. Reliability curve for GB+Graph on the test set (diagonal is perfect calibration)

For ecological mechanisms of top predictors, the dominance of lagged mean temperatures (tavg_mean_lag2 and tavg_mean_lag3) aligns with the established extrinsic incubation period of WNV. Higher temperatures accelerate viral replication within the mosquito; the 2-to-3-week lag represents the time required for a mosquito to become infectious and be detected in a trap.

Furthermore, the importance of dryness signals (rh_min and lagged precipitation) supportsthe"drought-amplification" hypothesis. In urban environments, relative dryness often stabilizes stagnant water in underground catch basins, concentrating mosquitoes and avian hosts. This creates an ecological bottleneck that accelerates transmission—a mechanism the GBDT model successfully recovers through the negative association between moisture lags and positivity risk.

6.5. Operational thresholding and confusion matrix

We select a decision threshold by maximizing F1 on validation (GB+Graph), then apply it to the test set. The resulting confusion matrix is shown in Figure 6; the full classification metrics appear in Table 5.

At this F1-tuned operating point, recall for the positive class is relatively high (0.697), capturing a meaningful share of risky weeks; precision is low (0.190), trading off specificity for sensitivity—a reasonable choice when the cost of missing risk is high relative to the cost of preventive action. Accuracy (0.675) is less informative under imbalance; the precision–recallspace (Figure 2) and calibration (Figure 4) better reflect usefulness.

For completeness, we include compact, faithful excerpts of the trap–week panel head and the test predictions (the full CSVs are in the archived ZIP).

6.6.Sensitivity analysis and ablation study

To assess the robustness of our spatial post-processing, we performed an ablation study on the graph construction (number of neighbors k) and the smoothing parameter α. The baseline case (α = 0) corresponds to the raw XGBoost output.

(i) Smoothing intensity (α): We varied α ∈ [0, 0.9] and observed that while AUROC remains relatively stable, the Brier score improves significantly as α increases from 0 to 0.7, indicating that spatial smoothing is the primary driver of the probability calibration gains reported in Table 2. At extreme values (α > 0.85), we observed a slight "over-smoothing" effect where local high-risk signals are excessively dampened by low-risk neighbors.

(ii) Intuitive impact of smoothing: Qualitatively,

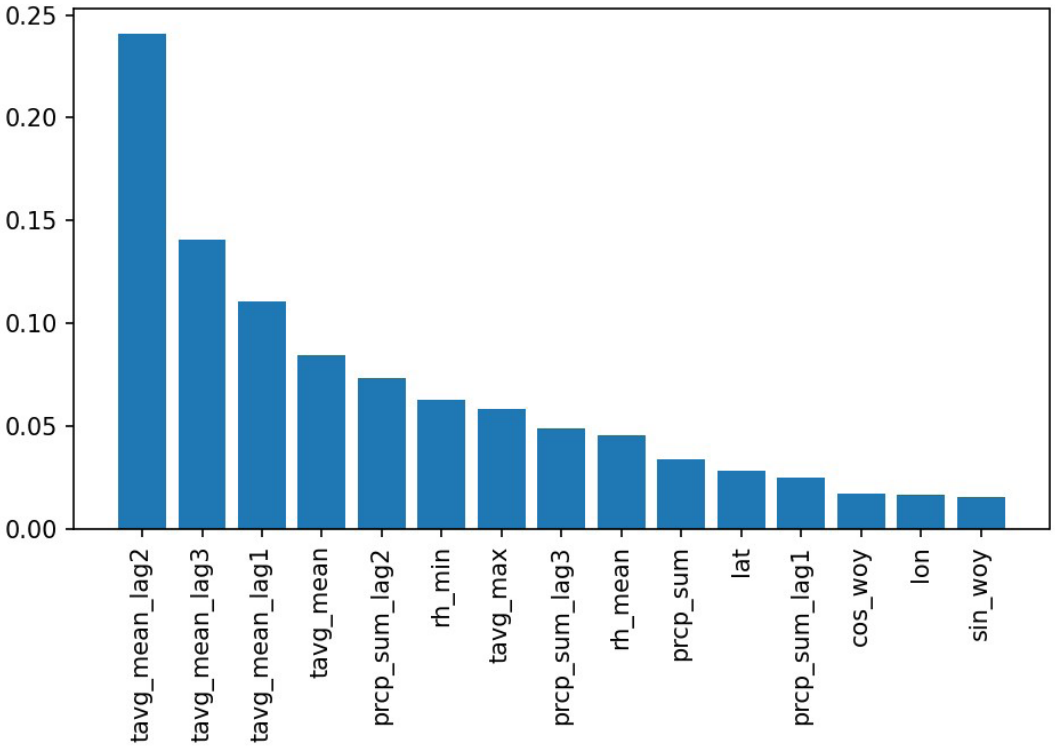

Figure 5. Gradient boosting feature importances

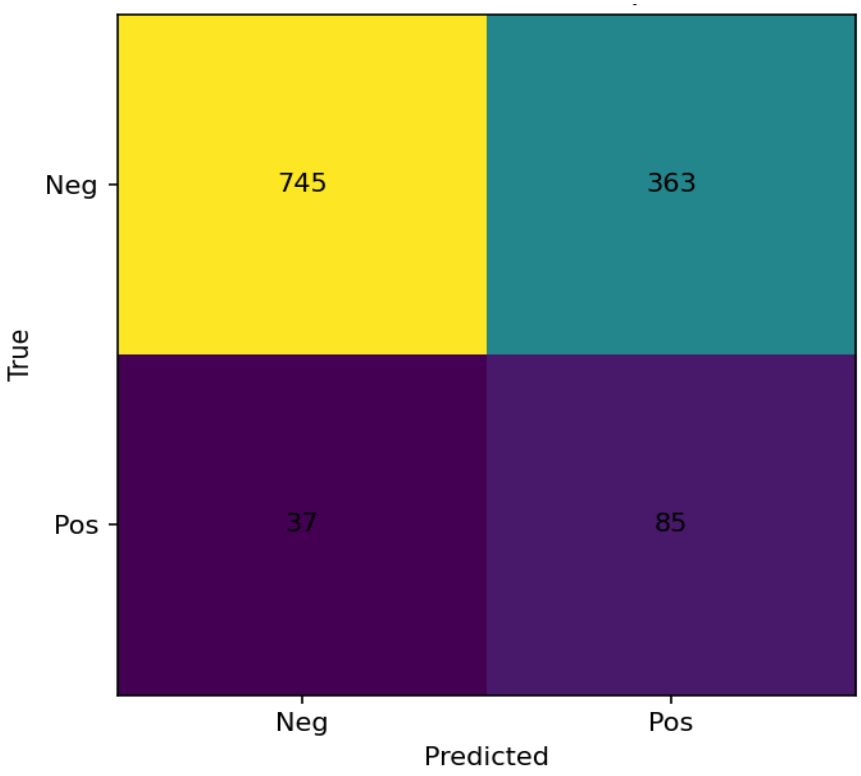

Figure 6. Confusion matrix at the F1-max threshold (GB + Graph, test set)
Abbreviations: Neg: Negative; Pos: Positive.

Table 3. Calibration table for GB+Graph on the test set (ten equal-frequency bins).

Bin	Mean $\hat{p}$	Observed Rate	Count
0	0.016	0.008	123
1	0.019	0.024	123
2	0.022	0.008	123
3	0.038	0.033	123
4	0.074	0.049	123
5	0.092	0.122	123
6	0.119	0.154	123
7	0.146	0.228	123
8	0.176	0.179	123
9	0.241	0.187	123

Table 4. Gradient boosting feature importances in descending order

Feature	Importance
tavg_mean_lag2	0.241
tavg_mean_lag3	0.141
tavg_mean_lag1	0.111
tavg_mean	0.085
prcp_sum_lag2	0.073
rh_min	0.063
tavg_max	0.058
prcp_sum_lag3	0.049
rh_mean	0.045
prcp_sum	0.034
lat	0.028
prcp_sum_lag1	0.025
cos_woy	0.017
lon	0.016
sin_woy	0.015

the Laplacian smoothing functions as a "spatial consensus" filter. Consider a scenario where a specific trap (Trap A) shows a high predicted probability based on local weather, but its five nearest neighbors report low probabilities. In the raw GB model, Trap A might trigger a false-positive alarm. After Laplacian smoothing, the low-risk "neighborhood consensus" pulls Trap A's probability down, treating the isolated

spike as potential measurement noise.

Conversely, if three adjacent traps show moderate risk, the smoothing reinforces this cluster, creating a more cohesive and reliable risk zone for municipal intervention. This process directly explains the dramatic improvement in the Brier score (Table 2) and reliability curve (Figure 4), as it moves "unconfident" isolated probabilities toward a more accurate regional mean.

(iii) Graph adjacency (kNN): We evaluated the impact of the neighborhood size k ∈ {3, 5, 8, 12}. Our results indicate that k = 5 provides a robust balance between capturing local spatial dependence and preventing "information bleed" from geographically distant traps. The model performance metrics were found to be stable for k between 3 and 8, suggesting that the framework is not overly sensitive to the exact graph construction method, provided basic spatial adjacency is maintained.

6.7. Takeaways

Table 6 shows that most trapweeks exhibit low winter temperatures and moderate humidity with no WNV detection, except for trap T1009, where a positive WNV case appears despite relatively cold conditions.

Four takeaways emerge:

(i) Spatially-aware smoothing improves probability quality: Despite similar AUROC/AUPRC to LogitEN, GB + Graph exhibits a dramatically lower Brier score (Table 2) and a reliability curve closer to the diagonal (Figure 4), yielding probabilities that are more actionable for thresholded decisions under resource constraints.

(ii) Robustness to spatial parameters: Sensitivity analysis confirms that the improvements in calibration are consistent across a range of smoothing parameters (α) and graph structures (k), identifying spatial regularization as a reliable enhancement for imbalanced surveillance data.

(iii) Seasonal and thermal drivers dominate, as expected in WNV ecology and embedded in the generator: lagged mean temperatures and seasonal harmonics rank highest (Table 4). Precipitation lags contribute a consistent (if smaller) signal, reflecting dryness effects on vector dynamics.

(iv) Sensitivity-oriented operating points are feasible: The F1-tuned thresholdcaptures a large fraction of positives (recall = 0.697) at the cost of lower precision, matchingoperational realities where missing high-risk weeks is more costly than investigating false alarms.

This analysis has several limitations. Results are from a synthetic generator designed to emulate a plausible spatio-seasonal structure. Real-world performance will depend on trap coverage, measurement quality, covariate completeness, and unmodeled drivers. The full repeat of this study with Chicago WNV pools plus Meteostat weather will further stress-test generalization and calibration.

7. Sustainability implications

As a premise,mosquito-borne risk is not merely an entomological phenomenon; it is a system-level outcome of how cities manage heat, water, vegetation, and waste. The results in Section 6—where recent heat load, relative dryness, and seasonality dominate risk and graph-aware smoothing yields well-calibrated probabilities—imply that

Table 5. Classification report for GB+Graph on the test set

Class/Avg	Precision	Recall	F1	Support
0	0.953	0.672	0.788	1108
1	0.190	0.697	0.298	122
accuracy	0.675	0.675	0.675	0.675
macro avg	0.571	0.685	0.543	1230
weighted avg	0.877	0.675	0.740	1230

Table 6.Trap–week panel (first 10 rows; selected columns)

trap_id	week_start	lat	lon	tavg_mean	tavg_max	prcp_sum	rh_mean	rh_min	wnv
T1000	2017-01-16	41.84	-87.78	0.06	2.84	1.81	47.73	30.87	0
T1001	2017-01-16	41.89	-87.53	2.90	7.34	0.22	39.80	30.00	0
T1002	2017-01-16	41.82	-87.61	3.98	8.06	3.32	47.53	33.96	0
T1003	2017-01-16	41.79	-87.76	3.51	7.63	4.47	42.59	30.00	0
T1004	2017-01-16	41.94	-87.62	2.34	6.68	3.71	48.06	30.45	0
T1005	2017-01-16	41.91	-87.73	2.65	10.89	2.79	40.42	30.00	0
T1006	2017-01-16	41.91	-87.58	1.02	4.03	3.91	46.36	35.05	0
T1007	2017-01-16	41.90	-87.74	1.03	5.17	3.40	46.19	30.00	0
T1008	2017-01-16	41.79	-87.65	1.36	5.72	5.59	52.06	39.80	0
T1009	2017-01-16	41.85	-87.51	-0.28	2.65	6.06	42.93	30.00	1

environmental design can shift the probability landscape that our model learns. Below, we translate model outputs into actionable levers for urban sustainability and vectorcontrol.

7.1. From probability to practice: A decision grammar

Because the GB + Graph model is comparatively well calibrated (low Brier; Figure 4), predicted probabilities pi,t can be mapped to *graded interventions* rather than a single hard threshold. Let $\tau_1 < \tau_2 < \tau_3$ be operating cutoffs chosen on validation (see also Section 5). We recommend a tiered playbook (Table 7) that prioritizesresource-stressed neighborhoods where environmental hazards and socioeconomic vulnerabilityintersect.

Because $\widehat{p_{i,t}}$ is probabilistic, agencies can also deploy *top*-K allocations (e.g., "treat the K traps with the largest $\hat{p}$ each week"), which optimizes field effort under fixed budgets.

7.2. Addressing confounding and bias

A primary limitation of surveillance-driven models is spatial sampling bias: traps are often denser in historically high-risk or higher-income areas. Our use of graph-Laplacian smoothing helps mitigate this by enforcing spatial coherence, ensuring that a "Critical" signal in one block elevates the concern for adjacent underserved blocks, even if they lack a physical trap. Furthermore, to address socioeconomic disparities, we recommend that sustainability co-actions (e.g., greening and cooling) be prioritized in census tracts where high $\widehat{p_{i,t}}$ coincides with high social vulnerability index (SVI) scores, preventing the systematic neglect of infrastructure-poor areas.

7.3. Design levers that change the odds

Our feature importance patterns (**Table 4**) suggest three high-yield levers whose benefits extend beyond vector control:

(i) Heat mitigation: Cool microclimates reduce thermal acceleration of the Culex–WNV cycle. Relevant

Table 7. Risk bands with associated operational and sustainability actions

Risk tier	Agency lead	Operational action	Sustainability co-action (practical implementation)
Caution	Public health	Enhanced surveillance & community outreach	Source reduction: Public Works teams clear alley debris, and public outreach campaigns encourage residential gutter cleaning.
Heightened	Vector control	Larviciding &trap densification	Water stewardship:Inspect municipal catch basins and deploy temporary "pocket cooling" (shade sails) in localized hotspots
Critical	Multi-agencytask force	Scheduled adulticide application windows	Infrastructure response: Conduct emergency desilting of bioswales and other drainage structures;high-frequency street watering to reduce microclimate heat load.

interventions include reflective or permeable pavements, high-albedo roofs, shade trees and vines,and cool corridors near historically hot blocks. These measures also provideco-benefits, including reduced heat stress, improved air quality, and energy savings.

(ii) Drainage and water stewardship: In our generator, dryness following wet periods is associated with elevated risk; managed drainage can shorten the lifespan of shallow, sun-warmed pools ideal for larvae. Interventions: bioswales, rain gardens with engineered soils, green alleys, downspout disconnections with safe dispersion, and maintenance of catch basins (de-silting). Co-benefits include flood mitigation and water quality improvements.

(iii) Vegetation structure: Vegetation affects shade, humidity, and water retention. Interventions: plant palettes and soil media that avoid persistent small water pockets; scheduled trimming to improve airflow where traps repeatedly flag high risk. Co-benefits include biodiversity support and urban aesthetics.

7.4. Practical implementation workflow

To bridge the gap between model output and field response, we suggest a weekly municipal cycle:

(i) Monday: Automated data ingestion and model inference. The system generates a ranked list of traps by probability $\widehat{p_{i,t}}$.

(ii) Tuesday: Decision-makers apply thresholds or top-K selection (e.g., treating only the 50 highest-risk blocks) based on the week's available budget.

(iii) Wednesday–Friday: Concurrent deployment. While Vector Control applies larvicide, Public Works crews are dispatched to "Critical" hotspots to remediate immediate structural drivers (e.g., clearing stagnant water in neglected drainage).

7.5. Targeting with spatial coherence

The Laplacian smoothing step exploits neighborhoodstructure: if several adjacent traps rise in risk simultaneously, the smoothed probabilities will reflecta coherent hotspot rather than isolated spikes. This supports microdistrict campaigns thatbundle actions over contiguous blocks, effectively counteracting the "blind spots" created byheterogeneous trap coverage.

7.6. Cost, equity, and accountability

Deployments must align with budgets and fairness:

(i) Cost-effectiveness: Use top-K targeting to maximize expected positives intercepted perstaff-hour.
(ii) Equity audits: Audit weekly coverage across neighborhoods; ensure that high-risk buthistorically underserved blocks—where neglected drainage or abandoned properties might amplify risk—are not systematically missed due to sampling bias.
(iii) Transparency: Calibrated forecasts (low Brier) enable honest communication of uncertainty,particularly when discussing risk in neighborhoods where surveillance data is sparse.

7.7. Ethical deployment and equity

The transition from model output to field action involves significant ethical responsibility.

(i) Avoiding stigmatization: Risk maps should be used for resource allocation (e.g., basin maintenance) rather than for penalizing specific neighborhoods. Care must be taken to communicate risk without stigmatizing high-risk tracts.
(ii) Algorithmic fairness: Because WNV risk often correlates with areas of historical disinvestment, there is a risk that models might reinforce existing disparities if they only prioritize high-resource surveillance areas. Our use of spatial smoothing and the recommendation to audit model performance against the SVI are intentional steps to ensure that high-risk, underserved blocks are not systematically overlooked.

7.8. Risks of misclassification and operational constraints

The primary risk in probabilistic operationalization is the trade-off between false negatives (missed outbreaks) and false positives (wasted resources).

(i) False Negatives: A missed signal in a "Caution" zone may lead to a localized outbreak. We mitigate this by using Laplacian smoothing, which ensures that even if one trap misses a signal, the risk in neighboring traps elevates the regional signal.
(ii) Resource Constraints: In high-risk weeks, the model may flag more "Critical" blocks than an agency has the capacity to treat. In such cases, the top-K approach ensures the highest-risk areas are prioritized, while lower-tier areas receive less labor-intensive "Source Reduction" outreach via digital channels.

7.9. Toward climate-ready surveillance

The most strategic sustainability investments are those that bend the risk curve before vector seasons begin. Because our pipeline is Python-native and reproducible, cities can run preseason counterfactuals: "What fraction of trap weeks would fall below τ_2 if (i) catch basins were cleared by week w_0, (ii) cool pavement were extended on m blocks, (iii) two rain gardens were added per hotspot?" In this way, planning becomes an integral part of early warning: using calibrated risk as a common metric to weigh trade-offs among cooling, drainage, and vector-control portfolios.

Overall,calibrated and spatially coherent risk estimates may enable agencies to move from reactive spraying towardmore anticipatory management—prioritizing where and when to intervene—and selecting a mix of sustainability and vector-control actions that address both the biological risk and theunderlying urban disparities.

7.10. Framework portability and generalizability

While our empirical results utilize synthetic panels emulating Chicago's urban morphology, the pipeline is architected for global portability across varying data environments and vector-borne diseases (e.g., Dengue or Zika). We distinguish between the broadly transferable logic and the context-specific parameters:

(i) Portable components:The Python-native modularity, graph-Laplacian smoothing algorithm, and the multitask evaluation protocol (AUROC vs. Brier) are universal. Any municipality with a Socrata-backed data portal and access to the Meteostat API can deploy the base pipeline with minimal code modification.
(ii) Context-specific adaptations: Local calibration is required for the temporal lags and decision thresholds (τ). For example, cities in tropical climates may observe shorter extrinsic incubation periods, necessitating 1-week lags rather than the 2–3-week lags identified for temperate Culex populations.
(iii) Ecological generalizability: The "drought-amplification" and "heat-loading" feature profiles are robust ecological drivers; however, researchers adapting the framework to different regions should evaluate whether additional covariates—such as the normalized difference vegetation index (NDVI) or localized hydrological flow models—should be stacked into the feature vector xi,t.

7.11. Limitationsandethics

Trap coverage varies across space and years; reporting and testing practices can change. Weather interpolation may misrepresent microclimates. We avoid causal claims and emphasize predictive use; data are public and de-identified, used under provider terms.

7.12. Confounding and non-intuitive effects

Meteorological drivers are strong predictors, but

their effects are potentially confounded by municipal interventions. For instance, a high temperature forecast might trigger proactive larvicide application by city agencies, which would suppress the observed positivity rate despite the weather being ecologically favorable for transmission. This "intervention bias" may explain why the importance of current-week precipitation is lower than historical lags. Additionally, urban microclimates—such as the urban heat island effect—can lead to localized variations in the extrinsic incubation period that are not fully captured by the regional weather stations used in this study, suggesting a need for more granular sensor integration in future iterations.

8. Discussion

Our results highlight the following themes:

(i) Calibration as the operational currency: Under rare positives, thresholds based solely on rank (ROC) can be misleading for action. The proposed Laplacian smoothing substantially improves Brier/reliability while leaving discrimination roughly unchanged, yielding probabilities that better support graded playbooks, expected-utility accounting, and transparent communication of uncertainty to residents and stakeholders.[4,5]

(ii) Ecological plausibility and levers: Feature profiles prioritize antecedent heat and relative dryness, consistent with known mosquito ecology and prior WNV modeling.[6] This alignment increases user trust and clarifies policy levers in urban design: microclimate cooling, rapid drainage of shallow warm pools, and vegetation management.[9,10] Because probabilities are calibrated, agencies can set tiered triggers and top-K allocations that balance sensitivity (missing risk) against crew-hours and chemical use.

(iii) Spatial coherence vs. local spikes: Smoothing amplifies coherent hotspots and dampens isolated spikes—useful for prioritization and for bundling actions over contiguous blocks—but risks masking acute hyper-local sources (construction debris, yard containers). We therefore recommend dual-view dashboards (raw and smoothed), learned adjacencies that incorporate hydrological or mobility pathways, and guardrail rules that escalate when either signal breaches a critical threshold.

(iv) Generalization and reproducibility: A strength of the framework is code parity between synthetic and real-data paths. This enables external reproduction on Chicago (and other cities with compatible endpoints) once credentials are supplied, addressing recurrent EWS concerns about transparency and auditability.[4] Still, performance will vary with trap coverage, testing practices, microclimates, and unmeasured environmental factors; the real-data replication is therefore a central next step.

(iv) Equity and accountability: Because arboviral risk and public services are not spatially uniform, we encourage fairness constraints during targeting (e.g., minimum coverage per ward above a probability band) and weekly reporting of coverage, false-alarm burden, and lift over random allocation. Calibrated probabilities make such accountability measures natural and legible.

(v) Operationalizing the EWS: We have detailed the practical hand-offs between departments. The strength of our framework lies in its multi-sectoral utility; while a Health Department uses the predicted probabilities for vector-control planning, the Department of Transportation or Planning uses the same values to justify "Green Alley" initiatives. This aligns WNV mitigation with broader climate resilience goals, making the EWS a core component of "Smart City" infrastructure.

These findings should nevertheless be interpreted in light of several important limitations. Among them are:

(i) Synthetic results: Current findings are reported on a synthetic, offline panel that emulates spatio-seasonal structure; real-world generalization can differ with surveillance practice and local ecology.

(ii) Measurement and coverage: Real trap networks are uneven over space and time; pool testing protocols, detection limits, and reporting cadences evolve.

(iii) Weather representation: Station- or reanalysis-based weather can miss microclimates; our humidity proxy and precipitation aggregation may smooth relevant extremes.

(iv) Confounding and causality: We offer predictive—not causal—inference; model coefficients and importances should not be interpreted as causal effects of sustainability interventions.

(iv) Spatial smoothing side-effects: Laplacian post-processing can oversmooth true local spikes if adjacency is mis-specified or traps are sparse.

In this paper, we presented a reproducible, API-driven framework for WNV early warningthat bridges the gap between ML and urban sustainability planning. By integratinggraph-Laplacian smoothing with GBDT, we achieved high-fidelity risk estimates thatprioritize probabilistic calibration—the metric most critical for graded municipal responses.Our results on a synthetic panel demonstrate that eco-climatic signals and spatial coherencecan successfully delineate WNV risk, even under the severe class imbalance typical of urbansurveillance.

9. Conclusion

We presented a fully reproducible, Python-native framework for trap–week risk forecasting that couples lagged meteorological drivers with graph-aware spatial smoothing. On a synthetic panel calibrated to Chicago-like seasonality and geography, GBDT with Laplacian post-smoothing delivered (i) ROC performance on par with elastic net and (ii) markedly better probabilistic calibration (lower Brier and improved reliability). Feature importance profiles emphasize antecedent heat and relative dryness, aligning with ecological prior knowledge. The pipeline is operationally practical for early-warning triage under resource constraints and maps directly onto sustainability levers (urban cooling, drainage, and habitat management).

The future work and deployment roadmapproposed here builds on the present study.Three stepsappear particularly important for real-world validation of the system:

(i) Pilot validation:We aim to partner with the Chicago Department of Public Healthor similar municipal agencies to run the system in "shadow mode" during a mosquito season.This will allow for the evaluation of model performance against real-time, non-synthetic trapdata and the assessment of the provided wnv_real_only.py script's operational utility.

(ii) Institutional partnerships:Successful deployment requires inter-departmental cooperation.We propose a pilot framework where the model's calibrated risk bands serve as a shareddecision-support tool between Public Health (for vector control) and Public Works (forsustainability-focused infrastructure maintenance, such as bioswale desilting).

(iii) Surveillance augmentation:Future research will explore the integration of non-traditional data streams, such as high-resolution satellite-derived vegetation indices (NDVI) and neighborhood-level SVI scores, to further refine the "equity-first" targeting of interventions.

Overall, the proposed framework serves as a template for "climate-ready" surveillance that links urban planning and public health through a transparent, data-driven EWS.

To support reproducibility,the exact real-data code path is included in the codebase and is intended to preserve methodological consistency between the synthetic and real-data workflowsonce a Socrata App Token is available,without requiring disclosure of credentials duringmanuscript submission.

Acknowledgments

The authors express their sincere gratitude to all contributors who supported this research.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Debashis Chatterjee
Formal analysis: All authors
Investigation: Sagnik Acharyya, Subhrajit Saha
Methodology: All authors
Supervision: Debashis Chatterjee
Writing–original draft: Subhrajit Saha, Debashis Chatterjee
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

All code required to reproduce the results—including synthetic data generation, model training, evaluation, and figure creation—is openly available at: https://github.com/debashisdotchatterjee/ML-Virus-Early-Warning-from-Weather-and-Urban-Sustainability-Signals. The repository contains both the synthetic mode and the real-data mode (requiring a Socrata token), along with a reproducible environment specification and output archiving scripts. Users can re-run the full workflow end-to-end using standard Python packages.

References

1. CDC (U.S. Centers for Disease Control and Prevention). Data and maps for West Nile. Posted August 19, 2025. Available from: https://www.cdc.gov/west-nile-virus/data-maps/index.html [Last accessed on 2025 Oct 25].
2. CDC (U.S. Centers for Disease Control and Prevention). West Nile Virus – Current year data (2025). Available from: https://www.cdc.gov/west-nile-virus/data-maps/current-year-data.html [Last accessed on 2025 Oct 25].
3. CDC (U.S. Centers for Disease Control and Prevention). West Nile Virus - Historic data (1999–2024). Available from: https://www.cdc.gov/west-nile-virus/data-maps/historic-data.html [Last accessed on 2025 Oct 25].

4. Li Z, Meng F, Wu B, *et al*. Reviewing the progress of infectious disease early warning systems and planning for the future. *BMC Public Health*. 2024;24(1):3080.

doi: 10.1186/s12889-024-20537-2

5. Pham CT, Nguyen HT, Le HHTC, *et al*. Challenges and strategies for the development and implementation of climate-informed early warning systems for vector-borne diseases: A systematic review. *Trop Med Int Health*. 2025;31(1):10-21.

doi: 10.1111/tmi.70045

6. Farooq Z, Rocklöv J, Wallin J, *et al*. Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers. *Lancet Reg Health Eur*. 2022;17:100370.

doi: 10.1016/j.lanepe.2022.100370

7. El-Sayed E, Eid MM, Abualigah L. Machine learning in public health forecasting and monitoring the zika virus. *Metaheuristic Optim Rev*. 2024;1(2):1-11.

doi: 10.54216/MOR.010201

8. Villanueva-Miranda I, Xiao G, Xie Y. Artificial intelligence in early warning systems for infectious disease surveillance: A systematic review. *Front Public Health*. 2025;13:1609615.

doi: 10.3389/fpubh.2025.1609615

9. Kache PA, Santos-Vega M, Stewart-Ibarra AM, Cook EM, Seto KC, Wasser MAD. Bridging landscape ecology and urban science to respond to the rising threat of mosquito-borne diseases. *Nat Ecol Evol*. 2022;6(11):1601-1616.

doi: 10.1038/s41559-022-01876-y

10. Zawarus P. Green infrastructure for mosquito control. In: *Architectural Factors for Infection and Disease Control*. Routledge; 2022:109-125.

doi: 10.4324/9781003214502-9

11. Erazo D, Grant L, Ghisbain G, *et al*. Contribution of climate change to the spatial expansion of West Nile virus in Europe. *Nat Commun*. 2024;15(1):1196.

doi: 10.1038/s41467-024-45290-3

12. Mandalapu A, Seong K, Jiao J. Evaluating urban fire vulnerability and accessibility to fire stations and hospitals in Austin, Texas. *PLOS Clim*. 2024;3(7):e0000448.

doi: 10.1371/journal.pclm.0000448

13. You J, Hu J, Jiang B. Non-stationarity and spatial spillover effects in artificial intelligence development: Implications for sustainable urban transformation. *Sustain Cities Soc*. 2025;131:106746.

doi: 10.1016/j.scs.2025.106746

14. Kazasidis O, Jacob J. Machine learning identifies straightforward early warning rules for human Puumala hantavirus outbreaks. *Sci Rep*. 2023;13(1):4882.

doi: 10.1038/s41598-023-30596-x

15. Malik I, Khattak WA. Enhancing urban health: machine learning applications in environmental management. *J Sustain Infrastruct Cities Soc*. 2023;8(1):1-21.

16. Zhang Y, Chen K, Weng Y, Chen Z, Zhang J, Hubbard R. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. *Expert Syst Appl*. 2022;198:116882.

doi: 10.1016/j.eswa.2022.116882

17. Sun H, Chen S, Li X, Cheng L, Luo Y, Xie L. Prediction and early warning model of mixed exposure to air pollution and meteorological factors on death of respiratory diseases based on machine learning. *Environ Sci Pollut Res*. 2023;30(18):53754-53766.

doi: 10.1007/s11356-023-26017-1

18. Keshava Murthy R. Early Detection and Prediction of Zoonotic Disease Events Using Event-Based Surveillance and Machine Learning. PhD dissertation. Washington State University; 2023.

doi: 10.7273/000005349

19. Treash J. Bridging urban planning and public health: Investigating the relationship between land use change and vector-borne disease risks in Ontario. Master's thesis. Queen's University; July 2022. Available from: https://queensu.scholaris.ca/server/api/core/bitstreams/1a3a60a2-ac95-4b1d-afed-207643712d55/content [Last accessed on 2022 Jul 01].

20. Alarcón JA. Exploring Relationships between Vector-Borne Diseases and Landscape Architecture: Aedes aegypti, Aedes albopictus and Landscape Architecture. Master's thesis. University of Washington. Available from: https://digital.lib.washington.edu/researchworks/items/c80d830b-e153-4f32-a023-5e5a5c0971ad [Last accessed on 2016 Sep 22].

21. Qiu Y. Scalable and Efficient Material Point Methods on Modern Computational Platforms. PhD dissertation. University of California; 2024. Available from: https://escholarship.org/uc/item/5t1436px [Last accessed on 2024 Jul 01].

22. Lesk C, Anderson W, Rigden A, *et al*. Compound heat and moisture extreme impacts on global crop yields under climate change. *Nat Rev Earth Environ*. 2022;3(12):872-889.

doi: 10.1038/s43017-022-00368-8

23. Liu C, Kershaw T, Eames ME, Coley DA. Future probabilistic hot summer years for overheating risk assessments. *Build Environ*. 2016;105:56-68.

doi: 10.1016/j.buildenv.2016.05.028

24. Chen X, Moraga P. Dengue forecasting and outbreak detection in brazil using LSTM: Integrating human mobility and climate factors. *medRxiv*. Preprint posted online March 3, 2025.

doi: 10.1101/2025.03.02.25323168

25. Verjee F. An assessment of the utility of GIS-based analysis to support the coordination of humanitarian assistance. PhD dissertation. The George Washington University; 2007. Available from: https://www.proquest.com/openview/63ac77f794afc21758eb2a7bae87d486/1?pq-origsite=gscholar&cbl=18750 [Last accessed on 2024 Jul 01].

26. Baldwin JW, Benmarhnia T, Ebi KL, Jay O, Lutsko NJ, Vanos JK. Humidity's Role in Heat-Related Health Outcomes: A Heated Debate. *Environ Health Perspect*. 2023;131(5):055001.

doi: 10.1289/ehp11807

27. Feng L, Lu J, Hu J, Irfan M, Wu K. Divergent carbon emission mitigation pathways toward sustainable development: Heterogeneous effects of the digital economy in urban centers versus boundary regions. *Sustain Cities Soc*. 2025;132:106808.

doi: 10.1016/j.scs.2025.106808

28. Tan L, Yang Z, Irfan M, Ding CJ, Hu M, Hu J. Toward low-carbon sustainable development: Exploring the impact of digital economy development and industrial restructuring. *Bus Strategy Environ*. 2023;33(3):2159-2172.

doi: 10.1002/bse.3584

29. Xue H, Cai M, Liu B, Di K, Hu J. Sustainable development through digital innovation: Unveiling the impact of big data comprehensive experimental zones on energy utilization efficiency. *Sustain Dev*. 2024;33(1):177-189.

doi: 10.1002/sd.3112

30. Ördek B, McGree J, Corry P, Spreafico C. Investigation of failures in rotational moulding using historical production dataset and machine learning. *Int J Adv Manuf Technol*. 2025;141(7-8):4291-4309.

doi: 10.1007/s00170-025-16925-6

31. Zhu H, Chen S, Irfan M, Hu M, Hu J. Exploring the role of the belt and road initiative in promoting sustainable and inclusive development. *Sustain Dev*. 2023;32(1):712-723.

doi: 10.1002/sd.2705

32. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. ACM Press; 2006:233-240.

doi: 10.1145/1143844.1143874

33. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. 2015;10(3):e0118432.

doi: 10.1371/journal.pone.0118432

34. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3.

doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

35. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*. ACM Press; 2005:625-632.

doi: 10.1145/1102351.1102430

36. Tashman LJ. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int J Forecast*. 2000;16(4):437-450.

doi: 10.1016/S0169-2070(00)00065-0

37. Afeld, Munoz C. sodapy 2.2.0. Posted August 31, 2022. Available from: https://pypi.org/project/sodapy/ [Last accessed on 2025 Oct 25].

38. SODA Developers. Getting started with the SODA Consumer API. Available from: https://dev.socrata.com/consumers/getting-started.html [Last accessed on 2025 Oct 25].

39. Meteostat Developers. Meteostat Python. Available from: https://dev.meteostat.net/python/ [Last accessed on 2025 Oct 25].

40. Meteostat Developers. API reference. Available from: https://dev.meteostat.net/python/api/ [Last accessed on 2025 Oct 25].

41. Meteostat Developers. Formats & units. Available from: https://dev.meteostat.net/formats.html [Last accessed on 2025 Oct 25].

42. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:785-794.

doi: 10.1145/2939672.2939785

43. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. *arXiv*. Preprint posted online 2017.

doi: 10.48550/arXiv.1705.07874

44. Smola AJ, Kondor R. Kernels and Regularization on Graphs. In: *Learning Theory and Kernel Machines (Lecture Notes in Computer Science)*. Springer; 2003:144-158.

doi: 10.1007/978-3-540-45167-9_12