Table of Contents

[CHAPTER 1: Research Overview and Previous Work 6](#_Toc193046613)

[1.1 Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreenin*g [14]* 7](#_Toc193046614)

[1.1 Metabolic Signature Discovery 9](#_Toc193046615)

[*CHAPTER 3:* Background on West Nile Virus and *Culex tarsalis* 13](#_Toc193046616)

[3.1 Introduction to Mosquito-Borne Diseases and West Nile Virus 13](#_Toc193046617)

[3.2 Environmental and climate drivers of WNV transmission 14](#_Toc193046618)

[3.3 The role of *Culex tarsalis* in WNV transmission and genetic adaptation 15](#_Toc193046619)

[3.4 Predictive Modeling of WNV transmission: challenges and limitations 15](#_Toc193046620)

[CHAPTER 4: Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis* 18](#_Toc193046621)

[4.1 Introduction 18](#_Toc193046622)

[4.2 Materials and Methods 19](#_Toc193046623)

[4.2.1 Sample Collection 19](#_Toc193046624)

[4.2.2 Genome Sequencing, Assembly, and Annotation 20](#_Toc193046625)

[4.2.3 RAD-Seq Library Preparation, Sequencing, and SNP Calling 20](#_Toc193046626)

[4.2.4 Climate Data Extraction 21](#_Toc193046627)

[4.2.5 Population Structure Analyses 23](#_Toc193046628)

[4.2.6 Identification of Genotype-Environment Associations 26](#_Toc193046629)

[4.2.7 Detecting SNPs Under Selection 28](#_Toc193046630)

[4.3 Results 28](#_Toc193046631)

[4.3.1 Genomic Analyses 28](#_Toc193046632)

[4.3.2 Population Structure 29](#_Toc193046633)

[4.3.3 Isolation by Distance (IBD) verse Isolation by Environment (IBE) 33](#_Toc193046634)

[4.3.4 Partial Redundancy Analysis 34](#_Toc193046635)

[4.3.5 Genotype-Environment Associations 35](#_Toc193046636)

[4.3.6 SNPs Under Selection 39](#_Toc193046637)

[4.4 Discussion and Conclusion 45](#_Toc193046638)

[CHAPTER 5: National West Nile Virus Prediction Model 48](#_Toc193046639)

[5.1 Data source 48](#_Toc193046640)

[5.1.1 Clinical Disease Surveillance Data 48](#_Toc193046641)

[5.1.2 Climate Data 49](#_Toc193046642)

[5.1.3 Land Cover Data 49](#_Toc193046643)

[5.1.4 Birds Data 50](#_Toc193046644)

[5.1.5 Demographic Data 50](#_Toc193046645)

[5.2 Methodology 50](#_Toc193046646)

[5.2.1 Data Preprocessing 50](#_Toc193046647)

[5.2.2 Standardization of data granularity 51](#_Toc193046648)

[5.2.3 Correlation Analysis using the Mantel Test 52](#_Toc193046649)

[5.2.4 Machine Learning / Statistical Models 52](#_Toc193046650)

[5.2.5 Training, Testing and Hyperparameter Tuning 53](#_Toc193046651)

[5.2.6 Evaluation Metrics 54](#_Toc193046652)

[5.3 Results 55](#_Toc193046653)

[5.3.1 National CDC Dataset 55](#_Toc193046654)

[5.3.2 Challenges in National Model 60](#_Toc193046655)

[CHAPTER 6: Regional West Nile Virus Prediction in California 66](#_Toc193046656)

[6.1 Data source 66](#_Toc193046657)

[6.1.1 Surveillance data 66](#_Toc193046658)

[6.1.2 Mosquitoes’ abundance data 69](#_Toc193046659)

[6.1.3 Data preprocessing 70](#_Toc193046660)

[6.1.4 Evaluation metrics 71](#_Toc193046661)

[6.2 Results 72](#_Toc193046662)

[6.2.1 Exploratory data analysis 72](#_Toc193046663)

[6.2.2 Iterative prediction framework 75](#_Toc193046664)

[6.2.3 Fixed training approach 77](#_Toc193046665)

[6.2.4 Hybrid autoregressive approach 79](#_Toc193046666)

[6.3 Discussion 83](#_Toc193046667)

[CHAPTER 7: Conclusion and Future Directions 86](#_Toc193046668)

[7.1 Summary of key findings 86](#_Toc193046669)

[7.2 Future directions 87](#_Toc193046670)

[7.3 Final remarks 87](#_Toc193046671)

[Reference 89](#_Toc193046672)

List of Figures

[Figure 1‑1: New ADAP-KDB spectral search algorithm with the prescreening search. 11](#_Toc193050227)

[Figure 1‑2 Execution times of the prescreening search algorithm. . 12](#_Toc193050228)

[Figure 1‑3 Metabolic Signatures Discover Pipeline 13](#_Toc193050229)

[Figure 1‑4 ST000085 PLS-DA results. 14](#_Toc193050230)

[Figure 1‑5 Upset plot between PLS-DA and ANOVA test. VIP cutoff is 1.3 and 1.5 in PLSDA; FDR threshold is 5% and 10% in ANOVA test. 15](#_Toc193050231)

[Figure 3-1: Correlations plot of each pair climate variables 26](#_Toc193050232)

[Figure 3-2 The cross-validation error plot of ADMIXTURE. Optimal number of Clusters (K) values ranging from 1 to 13. 27](#_Toc193050233)

[Figure 3-3 LFMM Workflow to Extract the Environmental Adaptive Candidate SNPs 30](#_Toc193050234)

[Figure 3-4 RDA Workflow to Extract the Environmental Adaptive Candidate SNPs 31](#_Toc193050235)

[Figure 3-5 Population Structure of Cx. tarsalis: (A) Floating pie charts of the admixture proportions in Cx. tarsalis populations sampled across the Western and Midwestern U.S and parts of Canada. Pie chart sizes are proportional to the sample size at each collection site. (B) ADMIXTURE results for K=4. Labels along the x-axis indicate sampling locations and colors correspond to the admixture proportion for each of the 4 clusters. (C) PCA results for the top 2 principal components, with points colored by the 4 geographic regions identified by ADMIXTURE. 33](#_Toc193050236)

[Figure 3-6 Scree plot for PCA 34](#_Toc193050237)

[Figure 3-7 Analysis of Molecular Variance (AMOVA). Gray bars indicate the expected distributions based on 999 random permutations, while black bars indicate the observed phi values. (A) Variation within populations, (B) Variation between populations within the same region, and (C) Variation between different regions. 35](#_Toc193050238)

[Figure 3-8 Isolation-by-Distance and Isolation-by-Environment. (A) Pairwise geographic distance versus genetic distance (FST) with best fit linear regression model (red line: ). (B) Pairwise environmental distance versus genetic distance with best-fit linear regression mode (red line: ). (C) Kernel density plot with best fit spline for geographic distance versus genetic distance. Areas of high, intermediate, and low density are represented by red, yellow, and blue colors, respectively. (D) Kernel density plot with best-fit spline for environmental distance versus genetic distance. Each point in panels A, B, C, and D represents one individual sample. 37](#_Toc193050239)

[Figure 3-9 Environmental Correlates of Genetic Variation in Cx. tarsalis Across Diverse North American Regions. Panels A and B display the relationship between environmental factors and the distribution of Cx. tarsalis, using RDA to illustrate how regional differences affect genetic variation. In these panels, the position of each circle (representing an individual mosquito) and color (indicating regional groupings from ADMIXTURE results) reflects their association with environmental variables, shown as purple vectors. The first plot (A) focuses on RDA1 and RDA2, the primary axes explaining the most variance, while the second (B) explores more subtle influences in RDA3 and RDA4. 41](#_Toc193050240)

[Figure 3-10 Correlation Heatmap between Environmental Variables and RDA Candidate SNPs. This heatmap illustrates the correlation between the environmental variables and candidate SNPs identified through the first four constrained axes of Redundancy Analysis (RDA). Each column on the x-axis represents a candidate SNP. For detailed SNP names and associated information on the 17 genes identified by SIFT4G, please refer to Supplemental Figure 3. 42](#_Toc193050241)

[Figure 3-11 BayeScan Results 43](#_Toc193050242)

[Figure 3-12 PCAdapt result. (Left) Q-Q plot to show the expected uniform distribution of the p-values. (Right) Histogram of p-values distribution 44](#_Toc193050243)

[Figure 3-13 Allele frequency distribution of selected candidate SNPs 44](#_Toc193050244)

[Figure 3-14 Upset Plot showing the overlap of genes annotated from candidate SNPs identified across four methods for detecting local adaptation and environmental associations in Cx. tarsalis: LFMM, RDA, PCAdapt, and BayeScan. 45](#_Toc193050245)

[Figure 4-1 National Annual Total Number of West Nile Virus (WNV) Human Cases (2004–2023) 58](#_Toc193050246)

[Figure 4-2 Kendall's Tau correlation Heatmap for National CDC Dataset: Annual Human Cases vs. Environmental Variables 60](#_Toc193050247)

[Figure 4-3 Comparative Analysis of SVM Models for WNV cases Across States 64](#_Toc193050248)

[Figure 4-4 WNV case distribution across the nation 66](#_Toc193050249)

[Figure 4-5 Comparative Analysis of SVM Models for WNV cases Across States Using Subsampling 67](#_Toc193050250)

[Figure 4-6 Annual human West Nile Virus (WNV) case counts (2004–2023) for selected U.S. states. (A) Annual case counts for individual states; (B) - (H) Annual case counts for California, Colorado, Illinois, Louisiana, Michigan, New York, Utah 68](#_Toc193050251)

[Figure 5-1 Annual Geographic Distribution of Human West Nile Virus Cases in California (2004–2023) 70](#_Toc193050252)

[Figure 5-2 Histogram plot of WNV cases different between CDPH and independently scraped data 71](#_Toc193050253)

[Figure 5-3 Heatmap of Monthly Mosquito Abundance (2004–2023) Across 13 California Counties, by Month. (A) Fresno; (B) Kern; (C) Los angeles; (D) Merced; (E) Orange; (F) Placer; (G) Riverside; (H) Sacramento; (I) San bernardino, (J) San joaquin; (K) Solano; (L) Stanislaus; (M) Tulare 73](#_Toc193050254)

[Figure 5-4 Kendall’s Correlation and P-value Heatmap between environmental variables and WNV disease count using CDPH-Official dataset 77](#_Toc193050255)

[Figure 5-5 Rolling statistics for number of WNV cases in California 78](#_Toc193050256)

[Figure 5-6 Iterative Forecasting Performance (Q2 Values) of SVM, RF, and HGBR Models 80](#_Toc193050257)

[Figure 5-7 Performance of SVM, RF, and HGBR When Trained on 2004–2018 and Tested on 2019–2023 81](#_Toc193050258)

[Figure 5-8 Example of construction of Lagged Temporal Features for WNV Prediction 82](#_Toc193050259)

[Figure 5-9 Result of Autoregression model using SVM, RF and HGBR 83](#_Toc193050260)

[Figure 5-10 Feature Importance Analysis Using Global SHAP Values for SVM, RF, and HGBR Models 84](#_Toc193050261)

[Figure 5-11 Impact of Temporal Lag on Random Forest Prediction Accuracy (Q² and MSE) 85](#_Toc193050262)

[Figure 5-12 Histogram of 95% Confidence Intervals of Q² from Bootstrapped Predictions for SVM, RF, and HGBR. SVM: (A) and (B); Random Forest: (C) and (D); Histogram-Based Gradient Boosting Regression: (E) and (F) 86](#_Toc193050263)

**Harnessing Computational Tools and Complex Biological Data for Environmental and Health Applications**

# Research Overview and Previous Work

The One Health framework is an interdisciplinary approach that recognizes the interconnectedness of human, animal, and environmental health[[1–4]](https://paperpile.com/c/wo71Nx/gyD4+7ggK+VjF3+fQ2V). It emphasizes the need for collaboration across medical, veterinary, and ecological sciences to address global health challenges. Originally developed to combat emerging infectious diseases, One Health has expanded to include fields such as antimicrobial resistance, food safety, and chronic disease epidemiology[[5]](https://paperpile.com/c/wo71Nx/jIgw). With increasing threats posed by zoonotic diseases, climate change, and environmental pollution, a holistic understanding of health at multiple scales is imperative.

A crucial component of the One Health paradigm is the utilization of big data to capture complex interactions across health domains[[6,7]](https://paperpile.com/c/wo71Nx/eUaY+vzEH). Advances in computational biology, epidemiology, and environmental monitoring have facilitated the integration of diverse datasets, including electronic health records, genomic sequencing, environmental pollution indices, and metabolomics data[[7–9]](https://paperpile.com/c/wo71Nx/Jja4+vzEH+voMO). The application of machine learning and high-performance computing has further enabled researchers to analyze large-scale datasets to detect health patterns, predict disease outbreaks, and assess environmental impacts on metabolic processes[[10–12]](https://paperpile.com/c/wo71Nx/nH7o+z6fr+zWoV).

Metabolomics, a rapidly growing field within One Health, provides a powerful tool for understanding metabolic interactions at the human-animal-environment interface. By examining metabolic signatures in biological samples, metabolomics can reveal biomarkers of exposure to environmental contaminants, dietary influences on health, and susceptibility to infectious diseases. The refinement of computational tools such as ADAP-KDB[[13,14]](https://paperpile.com/c/wo71Nx/RtKn+4BlE) enhances our ability to process and analyze mass spectrometry data, enabling more accurate identification of metabolic perturbations relevant to disease mechanisms.

By embedding the One Health approach into metabolomics research, we can improve our capacity to interpret metabolic changes that signal emerging health threats. This perspective fosters interdisciplinary collaborations among bioinformaticians, epidemiologists, ecologists, and healthcare professionals, ultimately contributing to more comprehensive public health strategies. The methodological advancements described in this chapter serve as a foundation for applying metabolomics within the One Health paradigm, paving the way for future studies that leverage metabolic insights for disease prediction, environmental health assessments, and translational medicine.

## Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreenin*g* [*[14]*](https://paperpile.com/c/wo71Nx/4BlE)

The original ADAP-KDB spectral search algorithm, which used a relational database for storing and querying spectral data, became increasingly slow with the growth of the spectral database. This slowdown was primarily due to the method used to calculate spectral similarity between a query spectrum and library spectra, which involved sqrt-cosine similarity calculations for each comparison. Each query spectrum needs to be compared with an increasingly large number of library spectra, a process that is computationally intensive. Additionally, using a general-purpose relational database like MySQL, while memory-efficient and cost-effective, may not be the fastest for spectral search, particularly when dealing with vast data. As the database grows, the time taken to match a query spectrum to all the library spectra in the database increases, leading to slower overall performance.

The new algorithm in ADAP-KDB is faster due to the implementation of a prescreening search step (Figure 1-1). This step allows the algorithm to quickly identify candidate spectra, reducing the need to calculate similarity scores for every library spectrum. For this speed improvement to be effective, three conditions must be met: (1) the pre-screening algorithm is much faster than the main search algorithm, (2) it returns a relatively small number of candidate spectra, and (3) the returned candidate spectra include the correct match to the query spectrum.

In this new algorithm, the process involves pre-calculating 𝑚/𝑧 values of the largest peaks in all library and query spectra. Then, the 𝑚/𝑧 value of the largest peak in the query spectrum is matched to 𝑚/𝑧 values of n largest peaks in the library spectra, and this process continues until the 𝑚/𝑧 value of the n-th largest peak in the query spectrum is matched to 𝑚/𝑧 values of m largest peaks in the library spectra, as illustrated in Figure 2 where (n = 8, m = 15). The library spectra are ranked based on the number of matched peaks. Candidate spectra are determined based on certain criteria, such as the number of matched peaks and a threshold value R. Spectral similarities between the query spectrum and candidate spectra are then calculated, and the candidate spectra with the highest scores are returned to the user.

Calendar

Description automatically generated

Figure ‑: New ADAP-KDB spectral search algorithm with the prescreening search. (A) m/z values of 8 largest peaks in the query spectrum are matched to m/z values of 15 largest peaks in every library spectrum; (B) all library spectra are ranked based on the number of matched m/z values, and top 50+ candidate spectra are returned by the prescreening search; (C) the similarity score is calculated for each candidate spectrum.

Figure 1-2 shows the comparison of execution time between the new library search algorithm and the original algorithm. The result shows that performance of the library search with pre-screening is about four-times faster for the low-mass-resolution spectra and very similar for the high-mass-resolution spectra. Moreover, the execution time stays about the same for all pairs of parameters n and m, while pairs (n = 4, m = 7), (n = 6, m = 11), and (n = 8, m = 15) demonstrate slightly better inclusion rate than pairs (n = 4, m = 4), (n = 6, m = 6), and (n = 8, m = 8), respectively. Based on these results, comparing eight largest peaks in the query spectrum to 15 largest peaks in the library spectra seems to be the optimal approach for the preliminary search. Selecting threshold R is based on the tradeoff between the execution time and the inclusion rate. Based on the comparison results, value R = 50 seems to be optimal for keeping high inclusion rate and low execution time for both low-mass-resolution and high-mass-resolution spectra. Therefore, (n = 8, m = 15 and R = 50) were selected as optimal for the prescreening search of the new spectral search algorithm.

Chart, calendar

Description automatically generated

Figure ‑ Execution times of the prescreening search algorithm, estimated for prescreening parameters (n, m, R), where n is the number of query spectrum peaks participating in the prescreening search, m the number of library spectrum peaks participating in the prescreening search, and R determines the number of candidate spectra returned by the prescreening search: (A,B) search against low-resolution spectra; (C,D) search against high-resolution spectra.

## Metabolic Signature Discovery

Figure 1-3 illustrates the workflow for discovering metabolic signatures. Raw data from public metabolomics studies undergo preprocessing using ADAPBIG, a software for processing untargeted mass spectrometry data [[15]](https://paperpile.com/c/wo71Nx/sZxJG). This step includes formatting the data for compatibility with machine learning models. Subsequently, the data undergo cleaning, imputation, scaling, and target creation using a dummy matrix. The processed data is then fed into Partial Least Squares Discriminant Analysis (PLS-DA), a supervised learning algorithm used for identifying metabolite patterns that differentiate sample groups. VIP scores are computed in PLS-DA to quantify the contribution of each metabolite to group separation, with a threshold typically set at 1. Metabolites surpassing this threshold are considered metabolic signatures. Finally, PLS-DA performance is validated through cross-validation, and the results are visualized.

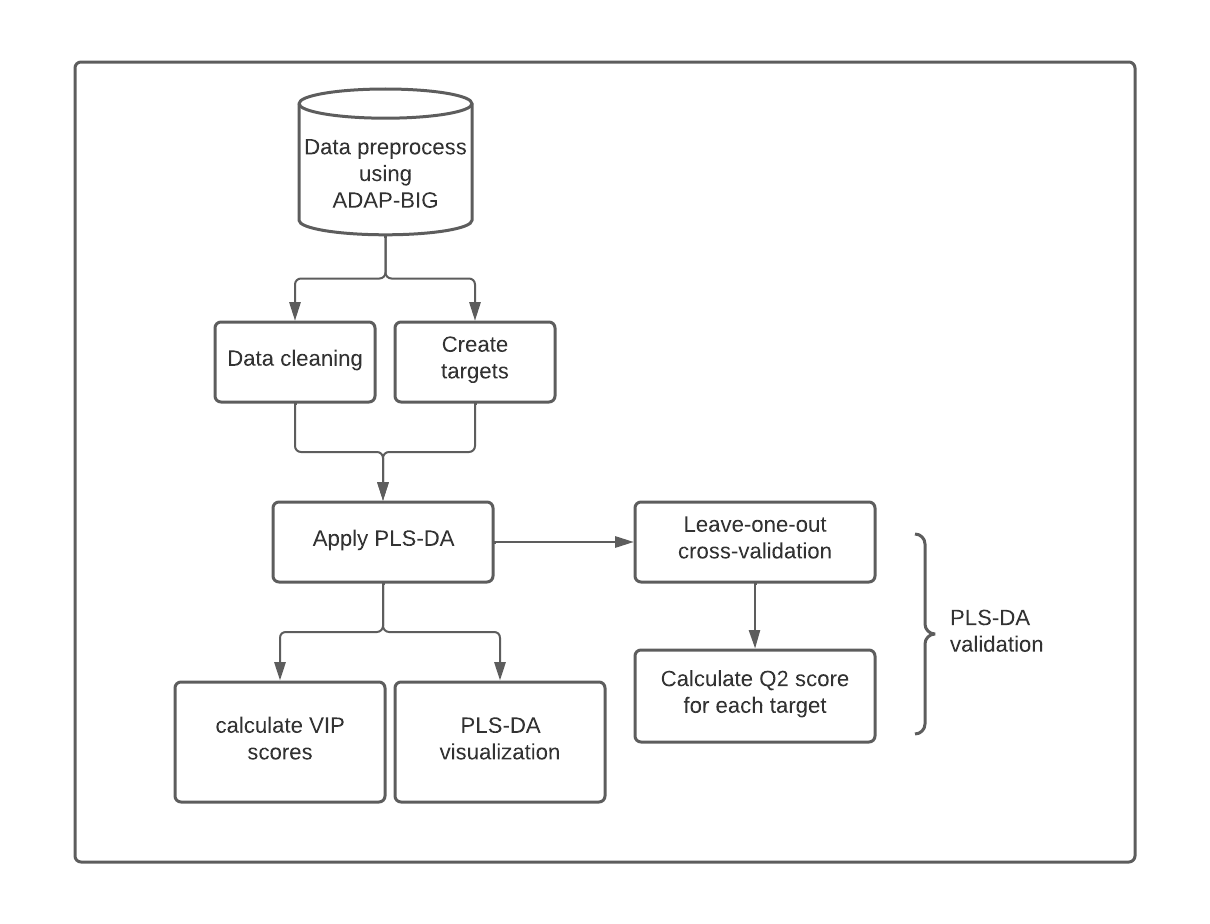


Figure ‑ Metabolic Signatures Discover Pipeline

Study ST000058 [[16]](https://paperpile.com/c/wo71Nx/TcLeQ) from the Metabolomics Workbench [[17]](https://paperpile.com/c/wo71Nx/ghC05) investigates alterations in metabolite levels associated with methionine stress sensitivity in cancer using GC TOF MS analysis. The study comprises seven groups, each containing four samples. Group 1 serves as the control, receiving methionine treatment, while groups 2 to 7 undergo homocysteine treatment for varying durations from 2 hours to 48 hours. This design allows for the examination of metabolite responses under different stress conditions, offering valuable insights into cancer metabolism.

Figure 1-4 presents the PLS-DA visualization of results across six different pairs of groups, demonstrating clear separation between each group. The predictive capability of the model is assessed through Q2 scores after leave-one-out cross-validation, with the highest Q2 observed between the control group and the 48-hour treatment group (0.76), and the lowest between the control group and the 8-hour treatment group (0.36).

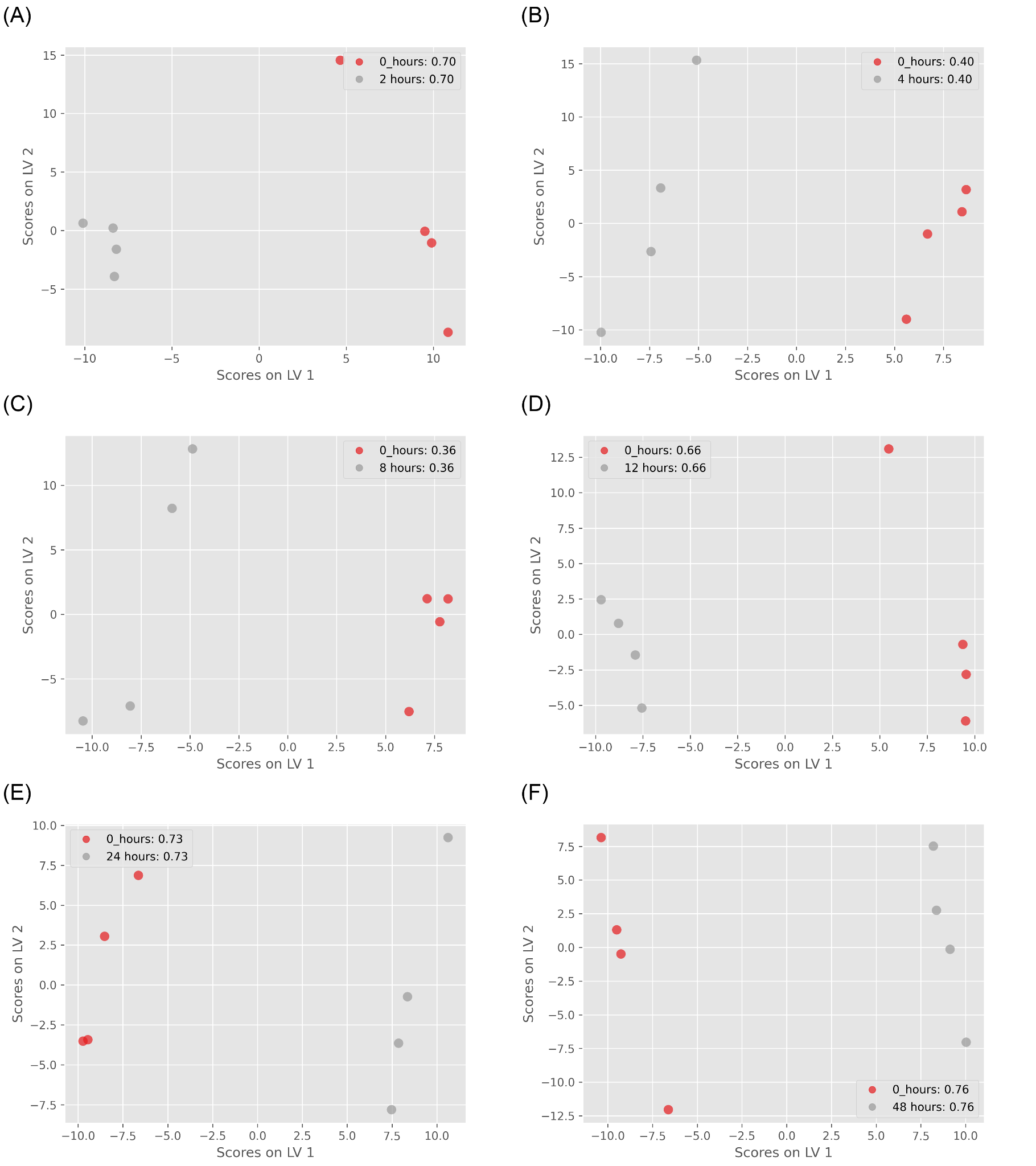


Figure ‑ ST000085 PLS-DA results. (A) control group vs. 2 hours treatment group (Q2=0.7); (B) control group vs. 4 hours treatment group (Q2=0.4); (C) control group vs. 8 hours treatment group (Q2=0.36); (D) control group vs. 12 hours treatment group (Q2=0.66); (E) control group vs. 24 hours treatment group (Q2=0.73); (F) control group vs. 48 hours treatment group (Q2=0.76);

To identify metabolic signatures, I conducted PLS-DA analysis on control groups and 48-hour treatment groups, utilizing VIP cutoffs of 1.3 and 1.5 to select candidate metabolites. Additionally, I performed ANOVA tests on the same group pair to compare multivariate and univariate algorithms, correcting the findings with 5% and 10% FDR thresholds. Figure 1-5 presents the comparison between PLS-DA and ANOVA results. A total of 26 metabolites were identified as candidates by both methods, with 19 uniquely identified by PLS-DA (using a VIP score cutoff of 1.3), and all metabolites from ANOVA were found in PLS-DA results.

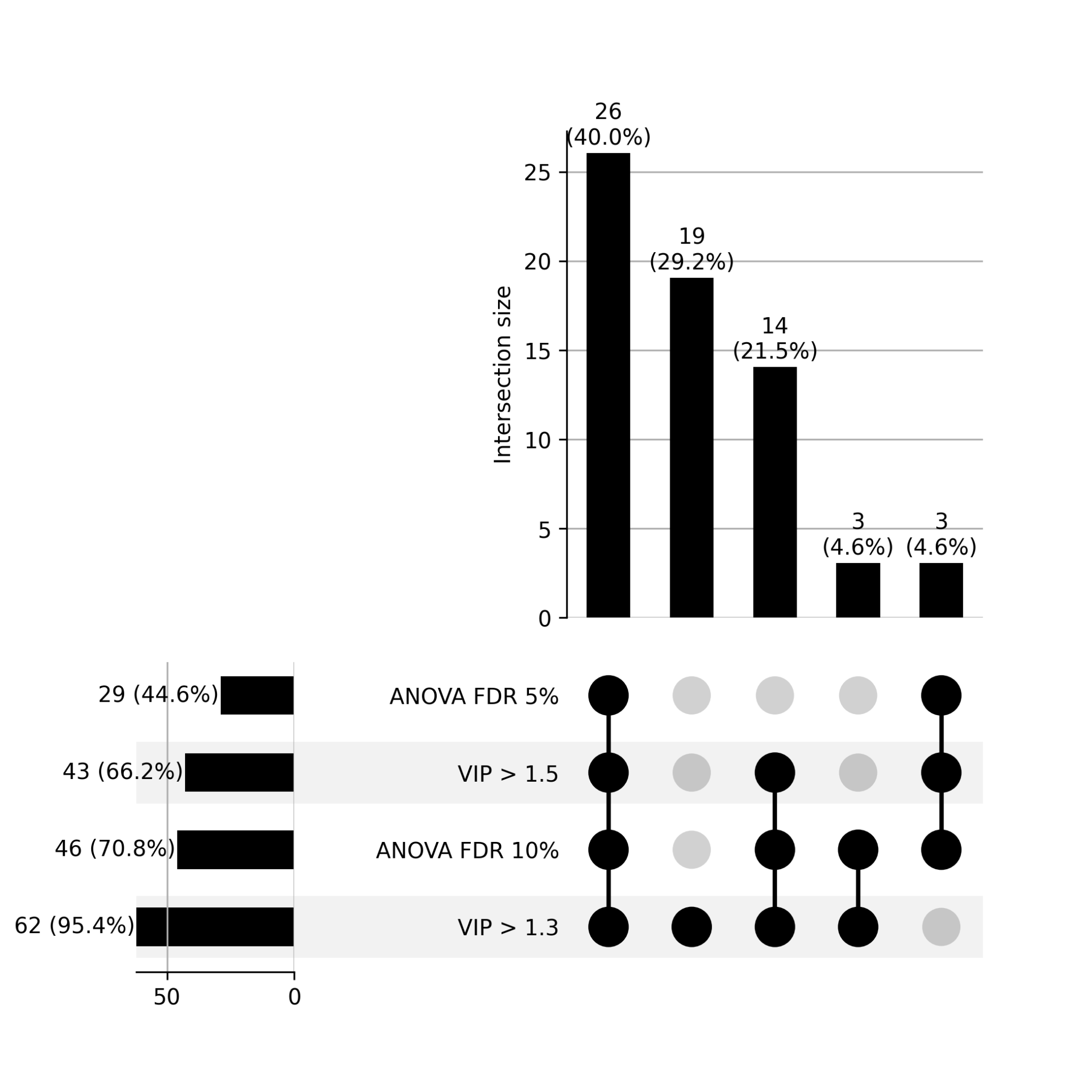


Figure ‑ Upset plot between PLS-DA and ANOVA test. VIP cutoff is 1.3 and 1.5 in PLSDA; FDR threshold is 5% and 10% in ANOVA test.

In refining the ADAP-KDB algorithm for metabolomics data analysis, I significantly enhanced the efficiency of processing mass spectrometry datasets, enabling rapid identification of metabolic signatures. This experience honed my skills in algorithm optimization, data preprocessing, and machine learning, equipping me with the tools necessary for tackling complex datasets in genetic and disease prediction research.

Applying these skills, I am poised to contribute to *Cx. tarsalis* population genetics by efficiently analyzing genetic variations and to West Nile Virus prediction by developing models that integrate multifaceted data. The foundation laid in computational data analysis and pattern recognition prepares me for advancing these projects, ultimately aiming to inform public health strategies.

# 

# Background on West Nile Virus and *Culex tarsalis*

## Introduction to Mosquito-Borne Diseases and West Nile Virus

Mosquito-borne diseases remain one of the most pressing public health challenges worldwide, affecting millions of people annually. These diseases are transmitted primarily by three major mosquito genera—Aedes, Anopheles, and Culex—each serving as a vector for different pathogens. Malaria, caused by Plasmodium parasites and spread by Anopheles mosquitoes, remains one of the deadliest mosquito-borne diseases globally, particularly in sub-Saharan Africa, South Asia, and parts of South America [[18–22]](https://paperpile.com/c/wo71Nx/xaAv+flLJ+q6xT+tDvQ+sYZ8). Meanwhile, Aedes-transmitted viruses, such as Dengue, Zika, and Chikungunya, have led to significant outbreaks in tropical and subtropical regions, with increasing cases in the southern United States due to climate and environmental changes[[22–30]](https://paperpile.com/c/wo71Nx/rrvh+HqFC+PZwc+cuuW+roTc+nk5a+wbg6+k4ed+sYZ8). Additionally, Culex mosquitoes, particularly *Culex tarsalis* and *Culex pipiens*, serve as primary vectors for West Nile Virus (WNV), St. Louis encephalitis (SLE), and Eastern equine encephalitis (EEE) in North America[[31–33]](https://paperpile.com/c/wo71Nx/gc0H+34a1+YabW).

Mosquito-borne diseases remain one of the most pressing public health challenges worldwide, with West Nile Virus (WNV) emerging as one of the most prevalent arboviral diseases in the United States. Since its initial detection in the United States in 1999 [[4]](https://paperpile.com/c/wo71Nx/fQ2V), WNV has spread across the entire United States in a few years, with over 59,000 reported cases and 2,900 fatalities by the end of 2023 based on the CDC report. Approximately 20% of those infected develop West Nile fever, with a critical 1% suffering from severe neurological illnesses[[34]](https://paperpile.com/c/wo71Nx/ZCoV).

Beyond its human health impact, WNV imposes substantial economic and healthcare burdens. Neuroinvasive WNV cases often result in extended hospitalizations, long-term rehabilitation, and permanent disability in some survivors[[35–38]](https://paperpile.com/c/wo71Nx/rJqE+GWnH+b2EX+g9VB). The direct costs of hospitalization and treatment for severe WNV infections, combined with lost productivity, pose a significant financial burden. Additionally, vector control programs—which include insecticide spraying, habitat modification, and public awareness campaigns—require extensive resources, with varying levels of success depending on environmental conditions and community engagement[[39–43]](https://paperpile.com/c/wo71Nx/f6XB+DWxa+KuO1+AWcw+Ekct).

Understanding WNV transmission dynamics is critical for developing effective surveillance systems and mitigation strategies. This requires an integrated approach that considers vector ecology, environmental factors, genetic adaptation, and predictive modeling techniques to enhance early warning systems and optimize public health interventions.

## Environmental and climate drivers of WNV transmission

WNV transmission dynamics are intricately influenced by a multitude of factors. Environmental conditions, including temperature, humidity, and precipitation, significantly impact mosquito populations and virus replication rates [[44,45]](https://paperpile.com/c/wo71Nx/6t59+2cTN). Moreover, geographical features such as land cover patterns and water bodies play crucial roles in shaping vector habitats and breeding grounds [[46]](https://paperpile.com/c/wo71Nx/scTxk). Population density and human behavior also contribute to WNV transmission, as urbanization and human activities can create favorable conditions for mosquito proliferation and contact with infected reservoir hosts [[47,48]](https://paperpile.com/c/wo71Nx/5Nwt8+QaxYv) .

* Temperature: Higher temperatures accelerate the extrinsic incubation period (EIP) of WNV within mosquitoes[[45,49–53]](https://paperpile.com/c/wo71Nx/QhMV+Ezhw+xKSt+JBjE+rkC0+2cTN), allowing the virus to replicate more rapidly and increasing transmission efficiency. Warmer conditions can also extend the mosquito breeding season and geographic range, potentially leading to earlier and more severe outbreaks.
* Precipitation and Drought: While heavy rainfall can create standing water habitats conducive to mosquito breeding, prolonged droughts have also been linked to WNV amplification [[45,50–56]](https://paperpile.com/c/wo71Nx/Ezhw+plCQ+ZTcg+xKSt+JBjE+rkC0+2cTN+rS46). Drought conditions concentrate bird populations around limited water sources, increasing the likelihood of virus transmission between mosquitoes and reservoir hosts.
* Urbanization and Land Use: The transformation of natural landscapes into urban areas creates environments favorable for WNV transmission[[51,57–59]](https://paperpile.com/c/wo71Nx/xKSt+XjYR+m77w+e0tz). Mosquitoes thrive in stormwater drains, artificial water containers, and irrigated landscapes, while urban heat islands can extend their activity periods and enhance viral transmission[[42,60,61]](https://paperpile.com/c/wo71Nx/bsf1+NSTO+AWcw).

Climate change is expected to further alter WNV transmission patterns by shifting mosquito habitats, modifying bird migration routes, and increasing extreme weather variability[[50,53,55,62–64]](https://paperpile.com/c/wo71Nx/Ezhw+ZTcg+rBN9+xZ28+rkC0+Cjfm). These environmental complexities underscore the necessity of predictive modeling frameworks that incorporate both climatic and ecological variables to anticipate future outbreak risks.

## The role of *Culex tarsalis* in WNV transmission and genetic adaptation

In the continental U.S., there are about 12 distinct mosquito species which can transmit diseases to humans, but not all of them have comprehensive genetic resources. *Cx. tarsalis* is a major vector for WNV in the United States [[65]](https://paperpile.com/c/wo71Nx/VU63r), and is a predominant vector of the disease in the most severely impacted states in the West and Midwest [[66]](https://paperpile.com/c/wo71Nx/J4uy3). Despite its significance, comprehensive genetic resources for *Cx. tarsalis* remains scarce, impeding our ability to elucidate its population dynamics and adaptability. Interestingly, our study of population genetics in *Cx. tarsalis* reveals a pattern of genetic differentiation that suggests a potential role for selection in addition to genetic drift. This pattern hints at environmental adaptations driving population divergence in *Cx. tarsalis*, suggesting that identifying the environmental factors and genetic determinants under selection is vital for predicting the spread of *Cx. tarsalis* and, by extension, WNV outbreaks.

## Predictive Modeling of WNV transmission: challenges and limitations

Recent studies have employed machine learning and statistical models to predict WNV transmission, leveraging historical case data, environmental conditions, and vector ecology. However, the predictive power of these models remains limited, as real-world outbreaks are influenced by complex, interdependent variables that are often difficult to quantify.

Key challenges in WNV prediction:

* Data limitations and surveillance bias

WNV surveillance data often suffer from underreporting and inconsistent case definitions, leading to gaps in historical records. Since these data are primarily obtained through government institutions responsible for epidemiological surveillance and meteorological monitoring, they are subject to limitations inherent in public health reporting systems[[67–69]](https://paperpile.com/c/wo71Nx/qVBu+ttyw+0Iof). Underreporting is a major concern, as mild or asymptomatic cases often go unreported when individuals do not seek medical attention, leading to systematic underestimation of disease prevalence[[68]](https://paperpile.com/c/wo71Nx/ttyw). Additionally, government health records frequently contain missing values and inconsistencies due to variations in data collection strategies across different institutions and states, complicating data-driven modeling efforts.

* Spatial and temporal complexity

**Spatial Heterogeneity**: WNV outbreaks exhibit strong spatial variability, influenced by differences in climate, mosquito species distribution, bird migration patterns, and human population density[[56,70–74]](https://paperpile.com/c/wo71Nx/VFss+Vt8m+4pvo+t1p4+0YRW+rS46). This means that a model trained in one region may not generalize well to another, as ecological and epidemiological factors driving transmission can vary significantly between locations. For example, an outbreak in a humid, urban environment with high mosquito control efforts may differ from one in a dry, rural area where vector populations fluctuate differently. These regional variations require models that account for local environmental and ecological factors to improve predictive accuracy.

**Seasonal Dynamics**: WNV transmission is highly seasonal, with peak activity typically occurring in late summer and early fall, depending on local temperature and precipitation patterns[[74–76]](https://paperpile.com/c/wo71Nx/mmNs+oxVV+0YRW). Mosquito population dynamics, virus replication rates, and bird-host availability change throughout the year, requiring models that can adapt dynamically to shifting environmental conditions[[77,78]](https://paperpile.com/c/wo71Nx/vtem+Vb7c). Additionally, extreme weather events such as heat waves or heavy rainfall can unexpectedly alter transmission patterns, making static models ineffective. Incorporating real-time environmental and meteorological data can improve model responsiveness to these seasonal variations.

This study applies a One Health approach to enhance West Nile Virus (WNV) prediction and control by investigating the environmental adaptation and transmission dynamics of *Cx. tarsalis*. Recognizing the connections within human, animal, and environmental health, it will analyze the population genetics of *Cx. tarsalis* to understand adaptive variations influencing vector competence and develop a machine learning-based model that integrates case trends, environmental factors, and mosquito population dynamics. Additionally, it will assess the impact of climate and habitat conditions on *Cx. tarsalis* distribution and identify key outbreak risk predictors. By incorporating ecological and epidemiological data, this research aims to provide actionable insights for public health, improving vector control strategies and resource allocation to mitigate WNV transmission within a One Health framework.

# Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis*

## Introduction

Diseases transmitted by insects are a serious and growing health concern for both humans and livestock. Numerous mosquito species are viral vectors, and the global distribution plus high adaptability of these mosquitoes contributes to the rapid spread and evolution of dangerous diseases across multiple continents [[79]](https://paperpile.com/c/wo71Nx/j4uom). Understanding how different species of mosquito have increased their ranges and adapted to new habitats in the past will be critical for predicting and managing their potential expansion in the future [[80,81]](https://paperpile.com/c/wo71Nx/hhdU6+q1K32).

The most abundant vector species found in the United States are derived from two genera: *Culex* and *Aedes*, each of which is comprised of both native and introduced species that occupy a variety of geographical ranges [[82]](https://paperpile.com/c/wo71Nx/kqVqG). *Culex tarsalis*, also known as the Western Encephalitis Mosquito, is a carrier of several forms of encephalitis that can infect humans as well as animals such as horses[[82,83]](https://paperpile.com/c/wo71Nx/kqVqG+H0Rzm). It is a known vector of West Nile Virus (WNV), Japanese Encephalitis Virus, St. Louis Encephalitis Virus, and Rift Valley Fever Virus [[84,85]](https://paperpile.com/c/wo71Nx/wE7w0+iRZhW). *Cx. tarsalis* is abundant in the western continental United States, where it is responsible for the majority of WNV cases in the most severely impacted states [[66,84]](https://paperpile.com/c/wo71Nx/wE7w0+J4uy3).

Despite the economic and health risks posed by this mosquito species, there is little known about its genetics. Previous studies using microsatellite markers have consistently uncovered a distinct pattern of population structure that does not entirely correlate with current geographical features or indicate strong isolation-by-distance [[86,87]](https://paperpile.com/c/wo71Nx/Y33Bl+booMa), but genetic results do support the hypothesis that *Cx. tarsalis* originated on the southwest coast of North America and has since undergone a range expansion to spread eastward [[88]](https://paperpile.com/c/wo71Nx/nL0aG). While significant adaptive changes must have occurred to allow populations to overwinter in order to cross the Rocky Mountains [[88,89]](https://paperpile.com/c/wo71Nx/nL0aG+lWmMT), the precise geographic and climatic variables driving divergence among populations are largely unknown.

Although the pattern of genetic differentiation observed in *Cx. tarsalis* could be the result of historical geographic divisions, it is also possible that local environmental adaptations may be driving some or all the population divergence in this species. Many mosquito species, including other *Culex* species, exhibit little to no population structure even after short time spans due to their large population sizes and short generation times [[90,91]](https://paperpile.com/c/wo71Nx/O4Np5+NmQJU), so the pattern observed in *Cx. tarsalis* is atypical and suggests a potential role for selection in addition to genetic drift. If the range of present-day populations of *Cx. tarsalis* is defined by adaptation to certain environmental factors, then identifying these factors as well as the genes and alleles under selection is essential to predicting whether or not *Cx. tarsalis* could continue to spread eastward and northward while being a more prevalent threat within the United States and possibly even other countries.

To advance our understanding of population structure and identify alleles linked to local adaptation in *Cx. tarsalis*, we first assembled and annotated a *de novo* reference genome and generated Restriction-Site Associated DNA sequencing (RAD-seq) data for over 300 individuals from 28 diverse geographic locations. We analyzed these RAD-seq markers through a comprehensive landscape genetics framework to explore how various environmental variables influence population differentiation and to identify alleles associated with adaptation to these conditions. By leveraging a broad spectrum of environmental variables, we assessed the adaptive responses of populations to their local environments, enabling the identification of critical genetic-environment associations. This approach reveals how specific climate variables and genetic variants underpin local adaptation strategies across 28 representative *Cx. tarsalis* collection sites. Our findings enrich our understanding of the complex interactions between genetics and environment, providing crucial insights into the ecological dynamics of this mosquito species.

## Materials and Methods

### Sample Collection

Individual mosquitoes were trapped and collected from 28 different locations across the United States and Canada as part of the North American Mosquito Project (NAMP) [[92]](https://paperpile.com/c/wo71Nx/cF5oS). All samples used in this study were collected in 2012 between the months of April and October.

### Genome Sequencing, Assembly, and Annotation

An F4 population was used to generate the reference genome assembly, and high molecular weight DNA was extracted and sequenced on a Pacific Biosciences (PacBio) RS II (University of Delaware). Thirty-five SMRTcells were generated. The resulting reads provided 76X coverage of the ~790Mb *Cx. tarsalis* genome, and were assembled with MECAT [[93]](https://paperpile.com/c/wo71Nx/QvI3a).

Gene annotation was completed by MAKER [[94]](https://paperpile.com/c/wo71Nx/qQ8xv) using EST and protein data from the *Culex quinquefasciatus* and *Aedes aegypti* mosquitoes. Sequences were downloaded from the NCBI Taxonomy database and both Trinotate and InterProScan were used for functional annotation of the MAKER predicted genes [[95,96]](https://paperpile.com/c/wo71Nx/Zxq2m+VjHQW). The annotated assembly was assessed for completeness and quality using BUSCO [[97]](https://paperpile.com/c/wo71Nx/y0T1W) and QUAST [[98]](https://paperpile.com/c/wo71Nx/wPrwE).

### RAD-Seq Library Preparation, Sequencing, and SNP Calling

DNA was extracted from individual mosquitoes and libraries were constructed for Restriction-site Associated DNA Sequencing (RAD-Seq) according to previously established protocols [[99]](https://paperpile.com/c/wo71Nx/AFmJ7). The SbfI enzyme was used to digest purified DNA, and individual samples were barcoded prior to Illumina sequencing. Raw sequencing reads were subsequently filtered to remove any reads with an uncalled base, an error in the restriction enzyme cut site, or with an average Phred quality score less than 20 over 15 consecutive nucleotides. Filtered reads were then de-multiplexed using the Stacks software package [[99,100]](https://paperpile.com/c/wo71Nx/AFmJ7+g58kA).

After de-multiplexing, raw reads from each individual were aligned to the draft assembly of the *Cx. tarsalis* genome using BWA MEM [[101]](https://paperpile.com/c/wo71Nx/6vy1P), and individuals with poor mapping rates (less than 50%) were excluded from subsequent analyses. The mapped reads for the remaining 378 samples were then merged using the Samtools pipeline [[102]](https://paperpile.com/c/wo71Nx/uHjqv) and SNPs were called using the GATK HaplotypeCaller [[103]](https://paperpile.com/c/wo71Nx/W1Ji2). The SNPs were filtered using VCFtools v0.1.12a [[103,104]](https://paperpile.com/c/wo71Nx/W1Ji2+zqxTr) to retain only sites with a minimum average individual read depth of 10X and a maximum of 20% missing data, resulting in a total of 457,387 sites. Individual samples were then filtered again to remove individuals with missing data at more than 20% of the remaining SNP sites, leaving 322 samples from 28 different locations for further analysis. The samples of the 28 different locations were grouped into 4 groups based on the ADMIXTURE results. Coding and noncoding SNP effects were predicted using SIFT4G [[105]](https://paperpile.com/c/wo71Nx/8BRG9).

### Climate Data Extraction

Climate data was extracted from the ERA5-Land monthly averaged dataset provided by the Copernicus Climate Change Service [[106]](https://paperpile.com/c/wo71Nx/fbT8b). The original dataset was characterized by a temporal resolution of 1 hour and a native spatial resolution of 9 km on a reduced Gaussian grid (TCo1279). To facilitate broader accessibility and suitability for diverse analyses, the data underwent regridding to a regular lat-lon grid with a finer resolution of 0.1x0.1 degrees.

To analyze environmental adaptation in *Cx. tarsalis*, we selected environmental factors based on their potential influence on the mosquito’s life cycle and their role in West Nile virus transmission dynamics[[56]](https://paperpile.com/c/wo71Nx/rS46). In this study, we initially extracted a total of 13 environmental variables for comprehensive analysis. These variables encompass a diverse range of climate parameters: 10m eastward wind, 10m northward wind, 2m temperature, evaporation from bare soil, leaf area index for high vegetation, leaf area index for low vegetation, water retention capacity of land, snowfall, surface net solar radiation, surface runoff, total evaporation, total precipitation, and volumetric soil water layer 1. Table 3-1 provides detailed descriptions of each variable.

Table 3- Description of environmental variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Short Name | Description |
| 10m eastward wind | eastward\_wind | The eastward component of wind speed at 10 meters above the ground. |
| 10m northward wind | northward\_wind | The northward component of wind speed at 10 meters above the ground. |
| 2m temperature | temperature | Air temperature measured at 2 meters above the ground. |
| evaporation from bare soil | evaporation | The amount of water evaporating directly from bare soil surfaces. |
| leaf area index for high vegetation | high\_vegetation | One-half of the total green leaf area per unit horizontal ground surface area for high vegetation type. |
| leaf area index for low vegetation | low\_vegetation | One-half of the total green leaf area per unit horizontal ground surface area for low vegetation type. |
| water retention capacity of land | water\_retention\_capacity | Amount of water in the vegetation canopy and/or in a thin layer on the soil. |
| snowfall | sf | The total amount of snow that falls over a specified period. |
| surface net solar radiation | ssr | The net amount of solar radiation reaching the earth's surface. |
| surface runoff | surface\_runoff | The amount of water that flows over the land surface after rainfall. |
| total evaporation | evaporation | The total amount of water evaporated from all sources, including soil and vegetation. |
| total precipitation | tp | Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface. |
| volumetric soil water layer 1 | swvl1 | Volume of water in soil layer 1 (0 - 7 cm) of the ECMWF Integrated Forecasting System. |

Prior to processing this complex dataset, steps were taken to minimize variable redundancy. This involved evaluating the variance for each variable across all samples, examining pairwise correlation coefficients, and assessing the distribution of each variable. Variables with a total variance across samples of zero, a pairwise correlation coefficient exceeding 0.70, or showing extreme distribution patterns were identified and eliminated. This refinement process led to the removal of 5 variables: surface net solar radiation, total evaporation, total precipitation, volumetric soil water layer 1, and snowfall. Consequently, the variable set was reduced to 8 environmental variables for all subsequent analyses (Figure 3-1).

A screenshot of a graph

Description automatically generated

Figure 3-: Correlations plot of each pair climate variables

The reduced dataset was processed using Python scripts utilizing the xarray and pandas libraries. These scripts calculated the averages for the 8 remaining climate variables for the year 2012. The data processing was specifically tailored to extract meaningful climate insights from NetCDF format files from the ERA5-Land monthly averaged dataset, with each variable's average computed based on its geographical location coordinates. This approach aimed to transform the extensive climate data into a format suitable for later Genome-Environment Association (GEA) [[107]](https://paperpile.com/c/wo71Nx/TUI6T) analyses.

### Population Structure Analyses

For population structure analyses, SNPs were further filtered using VCFtools v0.1.16 [[104]](https://paperpile.com/c/wo71Nx/zqxTr) to retain only bi-allelic sites with a maximum of 20% missing data per site and a minimum minor allele frequency (MAF) of 5% (Table 3-2).

Table 3- Filter on SNP file for each method in the paper

|  |  |
| --- | --- |
| Methods | Filter applied on original SNP file |
| Admixture | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| AMOVA | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| LFMM | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| RDA | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| Bayescan | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| PCAdapt | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |

The ADMIXTURE v1.3.0 program [[108]](https://paperpile.com/c/wo71Nx/lSxQQ) was subsequently employed on the remaining 17,239 loci and 322 individuals. Multiple runs of ADMIXTURE were conducted for K values ranging from 1 to 13, with each K value analyzed across 10 independent runs using different random number seeds in order to ensure convergence. The best K value was determined through a cross-validation process ranging from K=2 to K=13 (Figure 3-2). Subsequently, ancestry proportion bar plots were generated using Python's pandas library (plot.bar()), and floating pie charts displaying ancestry proportions over the U.S. map were plotted using the ggplot2 package in R [[109]](https://paperpile.com/c/wo71Nx/k1A5A) along with scatterpie() in the scatterpie package [[110]](https://paperpile.com/c/wo71Nx/3OwuH).

A line graph with dotted lines

Description automatically generated

Figure 3- The cross-validation error plot of ADMIXTURE. Optimal number of Clusters (K) values ranging from 1 to 13.

Complementing this, a Principal Component Analysis (PCA) was performed in R to aid in determining the pattern of genetic differentiation and the optimal K value by visualizing the dataset in a reduced dimensional space. As PCA requires no missing data, the missing genotype values for a given SNP were imputed using the most common genotype at each SNP across all individuals.

An Analysis of MOlecular VAriance (AMOVA) [[111]](https://paperpile.com/c/wo71Nx/bC957) was performed using the poppr package's poppr.amova() function in R [[112]](https://paperpile.com/c/wo71Nx/Ich55) [[111]](https://paperpile.com/c/wo71Nx/bC957) to detect population differentiation among regions inferred from ADMIXTURE results. Pairwise FST values between all populations were computed using the hierfstat package's genet.dist() function in R [[113]](https://paperpile.com/c/wo71Nx/WAQ2U), employing the Weir-Cockerham estimator [[114]](https://paperpile.com/c/wo71Nx/pgvwC). To assess whether there was a statistically significant level of population structure, a randomization test was conducted using the randtest() function from the ade4 package in R [[115]](https://paperpile.com/c/wo71Nx/jHh4r), employing 999 replicates.

To analyze isolation by distance (IBD) [[116]](https://paperpile.com/c/wo71Nx/fGiXu) and isolation by environment (IBE) [[117,118]](https://paperpile.com/c/wo71Nx/R1JLj+ETj2R) (Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014 Dec;23(23):5649-62. doi: 10.1111/mec.12938. Epub 2014 Oct 16. PMID: 25256562.) (Chang, CW., Fridman, E., Mascher, M. *et al.* Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant. *Heredity* 128, 107–119 (2022).) patterns within the 28 mosquito populations, genetic distances, derived from Weir & Cockerham FST estimations [[114]](https://paperpile.com/c/wo71Nx/pgvwC) based on SNP data, were compared with geographic and environmental distances computed based on latitude and longitude coordinates using the Haversine distance. Environmental distances were calculated in R using the dist() function, applying the Canberra method to emphasize relative differences in 8 non-negative environmental variables sourced from the Copernicus Climate Change Service for each geographic location. Scatterplots were created in R and linear regression models were fitted to each plot. Mantel tests [[119,120]](https://paperpile.com/c/wo71Nx/BLwKk+IDtbY) were used to assess the correlations between genetic distance and either geographic or environmental distances. The significance of these relationships was determined using 9999 permutations.

A mixed model was utilized to analyze the relationships between genetic distance, geographic distance, and environmental distance. To address potential collinearity issues that could create confounding results in the mixed model, variance inflation factor (VIF) assessments [[121]](https://paperpile.com/c/wo71Nx/TqaU7) were conducted to confirm low multicollinearity for each of the three distance matrices (Table 3-3).

Table 3- Variance Inflation Factor results for mixed model variables.

|  |  |
| --- | --- |
| Variables | VIF |
| genetic distance | 1.46 |
| geographic distance | 1.49 |
| Environmental distance | 1.04 |

To complement this analysis, a two-dimensional kernel density calculation was applied to visualize the concentration of data points within the scatterplots. To discern potential non-linear relationships between the distance matrices, a locally estimated scatterplot smoothing technique, implemented via the loess.smooth() function in R [[122]](https://paperpile.com/c/wo71Nx/xTOxg) , was employed. This nonlinear fit was compared against the linear model, with R² scores for both models calculated to assess and contrast their respective fits.

### Identification of Genotype-Environment Associations

To explore potential adaptive divergence in *Cx. tarsalis* populations, we employed two different approaches: a genome-environment association (GEA) [[107]](https://paperpile.com/c/wo71Nx/TUI6T) via latent factor mixed models (LFMM) [[123,124]](https://paperpile.com/c/wo71Nx/x6znI+2etUs)[[119,120]](https://paperpile.com/c/wo71Nx/BLwKk+IDtbY) and a redundancy analysis (RDA) [[125]](https://paperpile.com/c/wo71Nx/8mLeK) implemented in the vegan package in R [[126]](https://paperpile.com/c/wo71Nx/17wPy). For both analysis, SNPs were filtered to retain only bi-allelic sites with a maximum of 20% missing data per site and a minimum minor allele frequency (MAF) of 5%.

The LFMM approach (Figure 3-3) utilizes a univariate testing framework, modeling each SNP and environmental variable using the lfmm\_test() function from the LFMM package in R [[124]](https://paperpile.com/c/wo71Nx/2etUs). We initially conducted a Principal Component Analysis (PCA) on the environmental variables, focusing on the first principal component—a linear combination of the 8 climate variables—as the predictor in the LFMM. We also consider the inclusion of the second and third principal components as additional predictors in a subsequent LFMM, as detailed in the Supplemental Materials (Supplemental Figure 8 and Supplemental Table 15 - 17)

A diagram of a model

AI-generated content may be incorrect.

Figure 3- LFMM Workflow to Extract the Environmental Adaptive Candidate SNPs

In assessing the association between SNPs and environmental variables using the LFMM, we evaluated the Genomic Inflation Factor (GIF) to determine the model's efficacy in handling potential confounding factors. In our analysis, a Genomic Inflation Factor (GIF) value of 1.23 was used to adjust the p-values for potential inflation due to confounding variables, a process that is automatically incorporated within the output of the lfmm\_test() function in R [[124]](https://paperpile.com/c/wo71Nx/2etUs). Subsequently, we converted these GIF-adjusted p-values to q-values using the qvalue() function in R. This conversion is to refine the significance thresholds for individual tests, particularly under the framework of multiple hypothesis testing. It enhances the precision of FDR control, crucial in large-scale testing scenarios like ours. We then employed FDR control measures using these q-values, identifying candidate results falling below our predefined FDR threshold of 0.1.

The RDA operates as a multifaceted ordination technique, evaluating multiple loci concurrently with environmental variables (Figure 3-4). For this analysis, the significance of both the overall RDA model and its individual constrained axes was assessed. This assessment utilized the anova.cca() function from the vegan package in R [[126,127]](https://paperpile.com/c/wo71Nx/17wPy+Io1tp), facilitating a comprehensive examination of the null hypothesis (an absence of a linear relationship between SNP data and environmental variables). To select candidate SNPs for local adaptation, we identified SNPs that significantly deviated from the mean loadings, using a threshold of 3 standard deviations. Statistical significance was determined based on p-values below the threshold of 0.001.

A diagram of a system

AI-generated content may be incorrect.

Figure 3- RDA Workflow to Extract the Environmental Adaptive Candidate SNPs

To visualize results and identify relationships between RDA components and other factors, we utilized the vegan package in R [[126,127]](https://paperpile.com/c/wo71Nx/17wPy+Io1tp)to generate RDA tri-plots between each pair of the RDA components. To enhance the interpretability of our ordination plots, we employed symmetrical scaling [[127]](https://paperpile.com/c/wo71Nx/Io1tp). This scaling method adjusts the SNP and individual scores by the square root of the eigenvalues, providing a clearer representation of the relationships between variables and samples.

### Detecting SNPs Under Selection

To find which SNPs that were significantly associated with environmental variables also showed signatures of natural selection, we performed a Bayesian selection inference as implemented in BayeScan [[128–130]](https://paperpile.com/c/wo71Nx/HDQbj+9k7am+bZlDe). We also independently identified outliers using a PCA-based method implemented in the pcadapt R package [[131]](https://paperpile.com/c/wo71Nx/r3Tog), and then filtered for common candidate SNPs that were identified as significant across 4 different analyses: LFMM, RDA, BayeScan, and PCAdapt.

## Results

### Genomic Analyses

The final genome assembly contained 968,887,694 bases divided into 7,478 contigs. The N50 was 451,230 bp. The annotation included 43,905 predicted genes. The assembly quality, assessed using BUSCO version 5.1.3 with the Diptera\_odb10 lineage dataset (3285 BUSCOs), showed 88.2% complete BUSCOs (82.6% single-copy and 5.6% duplicated), 4.4% fragmented BUSCOs, and 7.4% missing BUSCOs. After aligning the RAD-seq reads and filtering for quality, there were 457,387 polymorphic sites identified across all populations.

### Population Structure

The ADMIXTURE analysis indicated a strong signature of population structure among the collected samples, with the optimal number of population assignments occurring at K=4 (Figure 3-2). The genetic clusters corresponded to four different broad geographic regions: (1) California/the West Coast, (2) the Southwest, (3) the Northwest, and (4) the Midwest (Figure 3-5).

The PCA results confirmed this pattern (Figure 3-5 and Figure 3-6), while also showing evidence of some sub-structure among the West Coast and Northwest populations (Figure 3-5C).

A map of the united states with different colored spots

Description automatically generated

Figure 3- Population Structure of Cx. tarsalis: (A) Floating pie charts of the admixture proportions in Cx. tarsalis populations sampled across the Western and Midwestern U.S and parts of Canada. Pie chart sizes are proportional to the sample size at each collection site. (B) ADMIXTURE results for K=4. Labels along the x-axis indicate sampling locations and colors correspond to the admixture proportion for each of the 4 clusters. (C) PCA results for the top 2 principal components, with points colored by the 4 geographic regions identified by ADMIXTURE.

A graph with a red line

Description automatically generated

Figure 3- Scree plot for PCA

The AMOVA results also indicated significant levels of population differentiation both between different populations within the same geographic region, and between different regions (Table 3-4 and Figure 3-7). The observed genetic variation within populations was significantly lower (p < 0.001) than expected (Figure 3-7A and Table 3-5), while variation between populations and between regions was significantly higher (p < 0.001) than would be expected by chance (Figure 3-7B and 3-7C, Table 3-5).

Table 3- Analysis of Molecular Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variation | Degree of Freedom (Df) | Sum of Squares | Mean Squares | % of Variation | Phi - statistics |
| Within samples | 290 | 410029.99 | 1413.9 | 78.86 | ɸ-populations-total = 0.21 |
| Between samples within regions | 28 | 51936.76 | 1854.88 | 2.54 | ɸ-populations-regions = 0.03 |
| Between regions | 3 | 85067.8 | 28355.93 | 18.6 | ɸ-regions-total = 0.19 |
| Total | 321 | 547034.55 | 1704.16 | 100 |  |

A group of graphs showing different types of data

AI-generated content may be incorrect.

Figure 3- Analysis of Molecular Variance (AMOVA). Gray bars indicate the expected distributions based on 999 random permutations, while black bars indicate the observed phi values. (A) Variation within populations, (B) Variation between populations within the same region, and (C) Variation between different regions.

Table 3- Permutation Tests of Observed and Expected Variation at different levels of population strata

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Monte-Carlo tests - permutation number: 999 | | | | |
| Test | Obs. | Std. Obs | Alter | P-value |
| Variations within populations | 1397.8 | -62.38 | less | 0.001 |
| Variations between populations within regions | 33.42 | 26.25 | greater | 0.001 |
| Variations between regions | 377.71 | 15.13 | greater | 0.001 |

### Isolation by Distance (IBD) verse Isolation by Environment (IBE)

Both geographic distance (IBD) and environmental distance (IBE) were tested for their relationships with genetic distance. The Mantel test revealed a strong and statistically significant correlation between genetic and geographic distances, with a Mantel statistic of 0.5661 (p < 0.001), indicating that geographic distance plays a notable role in genetic differentiation in this system. In contrast, the relationship between genetic and environmental distances was weak and non-significant, with a Mantel statistic of -0.05185 (p = 0.7741).

The mixed model analysis provided further insights. The full model, which incorporated both geographic and environmental factors, offered a more comprehensive explanation of genetic distance variations than models considering either factor alone (Table 3-6). Although the geographic distance model had a slightly lower Bayesian Information Criterion (BIC), the full model’s lower Akaike Information Criterion (AIC) suggests that combining geographic and environmental factors better captures the complexity of genetic differentiation. While geographic distance appears to have a stronger influence, these results suggest that environmental factors may contribute in more subtle or context-dependent ways to genetic variation. The kernel density and LOESS plots in Figure 3-8 highlight these relationships. Figure 3-8C shows a moderate to strong non-linear relationship between geographic and genetic distances, suggesting variability across different geographic ranges and the influence of complex factors beyond simple isolation by distance. The close alignment of the LOESS fit with the linear regression in Figure 3-8D suggests that most of the variation in the relationship between genetic and environmental distances can be captured by a simple linear model. However, the weak correlation (low R² of 0.0014) indicates that environmental factors may not exert a strong, direct influence on genetic variation at the global scale. This uniformity may mask subtle or context-specific interactions that are not evident in pairwise relationships.

A collage of different colored graphs

Description automatically generated

Figure 3- Isolation-by-Distance and Isolation-by-Environment. (A) Pairwise geographic distance versus genetic distance (FST) with best fit linear regression model (red line: ). (B) Pairwise environmental distance versus genetic distance with best-fit linear regression mode (red line: ). (C) Kernel density plot with best fit spline for geographic distance versus genetic distance. Areas of high, intermediate, and low density are represented by red, yellow, and blue colors, respectively. (D) Kernel density plot with best-fit spline for environmental distance versus genetic distance. Each point in panels A, B, C, and D represents one individual sample.

Table 3- Mixed model results for IBD and IBE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | AIC | BIC | k | AICc | AICcmin | BICew |
| Full | 695.801 | 715.476 | 5 | 695.963 | 0.526 | 0.137 |
| Distance | 696.061 | 711.801 | 4 | 696.169 | 0.474 | 0.863 |
| Environment | 1009.195 | 1024.934 | 4 | 1009.302 | 0.000 | 0.000 |

### Partial Redundancy Analysis

To further investigate the potential role of environmental factors, we performed partial redundancy analysis (partial-RDA) to separate and evaluate the individual contributions of geographic (IBD) and environmental (IBE) factors to genetic differentiation. Partial RDA results revealed that environmental factors, when controlling for geographic effects, explained 3.1% of the variance in genetic differentiation *( = 0.0309, p = 0.001).* In contrast, geographic factors, when controlling for environmental effects, explained a smaller yet statistically significant proportion of variance *(1.34%,  = 0.0134, p = 0.001).* These findings suggest that environmental factors may play a more prominent role in shaping genetic differentiation in the studied system, even when accounting for geographic structure.

Notably, among the environmental variables, low vegetation and evaporation showed the strongest influence on the first canonical axis (RDA1), highlighting their significant roles in shaping genetic variation (Table 3-7). Eastward wind contributed heavily to the second axis (RDA2), while water retention capacity and high vegetation were key drivers on the third and fourth axes (RDA3 and RDA4). These results suggest that specific environmental factors, particularly those related to vegetation and water dynamics, play a notable role in genetic differentiation, even after accounting for geographic effects.

Table 3- Biplot Scores for Constraining Variables in Partial RDA (Environment Controlled for Geography)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | RDA1 | RDA2 | RDA3 | RDA4 | RDA5 | RDA6 |
| eastward\_wind | -0.04295 | 0.6138 | -0.38221 | -0.068225 | 0.2848 | 0.4225 |
| northward\_wind | -0.01733 | -0.1283 | -0.60759 | 0.469732 | -0.1585 | -0.2478 |
| temperature | 0.25899 | -0.2221 | -0.03984 | 0.01674 | 0.1733 | -0.1052 |
| high\_vegetation | 0.07189 | 0.4416 | 0.1104 | 0.411215 | 0.4095 | -0.3103 |
| low\_vegetation | -0.78515 | -0.3226 | 0.0345 | -0.004408 | 0.1817 | -0.2035 |
| water\_retention\_capacity | -0.50671 | -0.2401 | 0.33451 | 0.665514 | 0.0183 | 0.1856 |
| surface\_runoff | -0.37392 | -0.4824 | 0.22363 | 0.196856 | 0.353 | 0.256 |
| evaporation | 0.56079 | 0.3324 | -0.01991 | -0.094839 | -0.2216 | 0.1812 |

### Genotype-Environment Associations

In the PCA of the 8 environmental variables, described in the Section 3.2, the first component explained 34.75% of the variation in the predictors, with the second component accounting for 21.42% of the variance, and the third component accounting for 16.73%, for a total cumulative explained variance of 72.9% (Table 3-8 and Table 3-9). Using the first PC as the predictor in the latent factor mixed model (LFMM), 92 candidate SNPs were identified as being significantly associated with the environment (FDR < 0.1) (Supplemental Table 5).

Table 3- PCA results of environmental variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Eigenvalue | 2.4326 | 1.4994 | 1.1709 | 0.69002 | 0.57216 | 0.457 | 0.17785 |
| Proportion Explained | 0.3475 | 0.2142 | 0.1673 | 0.09857 | 0.08174 | 0.06529 | 0.02541 |
| Cumulative Proportion | 0.3475 | 0.5617 | 0.729 | 0.82757 | 0.90931 | 0.97459 | 1 |

Table 3- Correlations between the PC axis and environmental variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Eastward wind | 0.323 | -0.377 | 0.369 | -0.399 | -0.142 | 0.303 | 0.529 |
| Northward wind | -0.164 | 0.374 | -0.339 | -0.805 | 0.079 | -0.183 | 0.157 |
| Temperature | 0.124 | 0.727 | 0.04 | 0.224 | -0.101 | 0.19 | 0.168 |
| High vegetation | 0.213 | 0.346 | 0.602 | -0.233 | 0.138 | 0.282 | -0.415 |
| Low vegetation | -0.4 | 0.137 | 0.373 | 0.039 | -0.718 | -0.29 | 0.189 |
| Water retention capacity | -0.414 | -0.11 | 0.472 | -0.102 | 0.43 | -0.391 | -0.127 |
| Surface runoff | -0.496 | 0.103 | 0.063 | 0.199 | 0.407 | 0.406 | 0.517 |
| Evaporation | 0.481 | 0.165 | 0.14 | 0.189 | 0.28 | -0.598 | 0.42 |

The redundancy analysis (RDA) model revealed that environmental variables explain 10.42% of the genetic variance (constrained), while 89.58% of the variance remains unexplained. This unexplained variance is presumably influenced by geographical distance, as supported by the results of the IBD/IBE and partial RDA tests. Nevertheless, the environmental contribution to genetic variance offers valuable insights into adaptive processes. Even though the majority of the variance was not explained by environmental factors, both the overall RDA model and the first 4 axes still demonstrated statistical significance, indicating that there were substantive associations within the data, which suggests the presence of meaningful genotype-environment associations (Table 3-10). A total of 822 SNP candidates were detected as significantly associated with at least one of the first four RDA loadings (Supplemental Table 6). Among these, 658 were found within genes. Additionally, out of all the genes identified by RDA containing the 658 SNPs, 32 genes were also found to be significantly associated with the environment in the LFMM.

Table 3- Significant test results on each constrained axis of RDA

|  | Df | Variance | Proportion | F | Pr(>F) | Significance |
| --- | --- | --- | --- | --- | --- | --- |
| Full Model | 8 | 1796.3 | 10.42% | 4.5511 | 0.001 | \*\*\* |
| RDA1 | 1 | 894.1 | 5.19% | 18.1217 | 0.001 | \*\*\* |
| RDA2 | 1 | 469.8 | 2.73% | 9.5231 | 0.001 | \*\*\* |
| RDA3 | 1 | 158.7 | 0.92% | 3.2158 | 0.001 | \*\*\* |
| RDA4 | 1 | 69.8 | 0.40% | 1.4154 | 0.001 | \*\*\* |
| RDA5 | 1 | 56.8 | 0.33% | 1.1517 | 0.094 | . |
| RDA6 | 1 | 51.1 | 0.30% | 1.0361 | 0.591 |  |
| RDA7 | 1 | 48.6 | 0.28% | 0.9855 | 0.901 |  |
| RDA8 | 1 | 47.3 | 0.27% | 0.9592 | 0.901 |  |
| Constrained axis Residual | 313 | 15442.7 |  |  |  |  |

Significant Code: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01: ‘\*’ 0.05 “.” 0.1 ‘ ’ 1

Figure 3-9A captures the distribution of mosquito populations across the RDA 1 and RDA 2 axes, which together explained 76% of the environmentally influenced genetic variance (corresponding to 5.84% of the total genetic variance). The clear regional clustering depicted within this biplot aligns with the four geographic regions (Midwest, Northwest, Southwest, and West Coast), underscoring a significant regional influence on the portion of genetic variation shaped by environmental factors. Specifically, the alignment of populations with vectors for the temperature and evaporation signifies the role of temperature and humidity in this context. Figure 3-9B further explores the subtler environmental gradients within the RDA 3 and RDA 4 axes, which together explain approximately 13% of the environmentally responsive genetic variance (equating to roughly 1.56% of the total genetic variance). Here, the distributions suggest a more intricate interaction of genetic variance with environmental variables like the surface runoff and leaf area index for high vegetation, which could reflect micro-environmental adaptations.

A screenshot of a graph

Description automatically generated

Figure 3- Environmental Correlates of Genetic Variation in Cx. tarsalis Across Diverse North American Regions. Panels A and B display the relationship between environmental factors and the distribution of Cx. tarsalis, using RDA to illustrate how regional differences affect genetic variation. In these panels, the position of each circle (representing an individual mosquito) and color (indicating regional groupings from ADMIXTURE results) reflects their association with environmental variables, shown as purple vectors. The first plot (A) focuses on RDA1 and RDA2, the primary axes explaining the most variance, while the second (B) explores more subtle influences in RDA3 and RDA4.

As in Figure 3-10 and Supplemental Table 7, the majority of the candidate SNPs in the RDA1 group are negatively correlated to evaporation, temperature and high vegetation. As shown in Figure 3-9A, the first and fourth quadrants likely represent eastern populations, while the second and third quadrants represent western populations. This east-west gradient is evident among the RDA1 SNPs group. Similarly, the RDA2 group candidate SNPs are more negatively correlated with low vegetation index and water retention capacity but positively correlated with temperature.

This analysis reveals that most significantly associated SNPs are correlated with more than one environmental variable and often cluster together in distinct patterns (Figure 3-10). Although this complex interaction represents a smaller slice of the total genetic variance, it is critical for a comprehensive understanding of genetic-environment relationships.

A colorful lines on a screen

Description automatically generated with medium confidence

Figure 3- Correlation Heatmap between Environmental Variables and RDA Candidate SNPs. This heatmap illustrates the correlation between the environmental variables and candidate SNPs identified through the first four constrained axes of Redundancy Analysis (RDA). Each column on the x-axis represents a candidate SNP. For detailed SNP names and associated information on the 17 genes identified by SIFT4G, please refer to Supplemental Figure 3.

This suite of RDA plots collectively reveals the environmental portion of genetic variance within *Cx. tarsalis* populations. While most of the genetic variation correlates with geographic distribution, the environmental variance captured here affords a critical perspective on the selective forces at play, contributing to the broader evolutionary narrative of this species.

Overall, our findings underscore the intricate connections between geographic locations, environmental factors, and genetic variations, confirming that in addition to genetic drift and IBD, there is a profound influence of environmental variables on genetic variation in *Cx. tarsalis*.

### SNPs Under Selection

The BayeScan outlier analysis, performed with a Q-value threshold of 0.05 to identify loci under selection, revealed 1,836 loci potentially under diversifying selection, 10,166 potentially under balancing selection, and 5,237 neutral loci (Figure 3-11 and Supplemental Table 8). Of the 1,836 loci putatively undergoing selection for local adaptation, 1501 of these are found within 824 genes. Of these 824 genes, 24 also overlapped with genes that were identified by both the LFMM and RDA as significantly associated with environmental variables.

A graph showing a bullet

Description automatically generated

Figure 3- BayeScan Results

We also used PCAdapt independently to identify outlier loci with notable allele frequency differences across populations, potentially resulting from natural selection. The Q-Q plot in Figure 3-12 (left plot) shows that most p-values follow the expected uniform distribution. However, the smallest p-values deviate significantly from expectation, indicating the presence of outliers. Similarly, the histogram of p-values (Figure 3-12 right plot) confirms this pattern, with most p-values aligning with a uniform distribution but an excess of small p-values highlighting the outliers. These outliers include 173 SNPs in the top 1% of extreme p-values. (Supplemental Table 9).

A graph with a line

Description automatically generatedA graph of a number of columns

Description automatically generated with medium confidence

Figure 3- PCAdapt result. (Left) Q-Q plot to show the expected uniform distribution of the p-values. (Right) Histogram of p-values distribution

In total, 53 candidate local adaptation SNPs were identified (Table 3.13). Some candidates show a strong East-West gradient in terms of allele frequencies (Figure 3-13A and B), while others indicate that the alternate allele is present only in one or a few populations (Figure 3-13C and D).

A group of graphs with different colored bars

Description automatically generated

Figure 3- Allele frequency distribution of selected candidate SNPs

Looking across all four analyses examining significant environmental associations and evidence of natural selection, we identified 20 common genes containing candidate SNPs that were consistently significant in each instance (Figure 3-14 and Supplemental Table 11). Given BayeScan’s susceptibility to potential false positives, we included PCAdapt to provide additional validation and strengthen the robustness of our findings.A Chi-Square test (Table 3-11 and Table 3-12) indicated a significant association between the candidacy of SNPs and their location being within genes rather than intergenic (X-squared = 75.842, df = 1, p-value < 2.2e-16). This suggests that SNPs identified as candidates are more likely to be located within genes than would be expected by chance.

A graph of different sizes and numbers

Description automatically generated with medium confidence

Figure 3- Upset Plot showing the overlap of genes annotated from candidate SNPs identified across four methods for detecting local adaptation and environmental associations in Cx. tarsalis: LFMM, RDA, PCAdapt, and BayeScan. The bar heights indicate the number of genes annotated based on candidate SNPs in each unique or shared category, with percentages relative to the total genes annotated. Black circles below the bars represent intersections of methods, where connected lines indicate the methods contributing to the overlap. For example, the tallest bar corresponds to 438 genes uniquely identified by BayeScan (47.9%), with no overlap with other methods. Individual horizontal bars represent the total number of genes identified by each method: LFMM\_PC1 (44, 4.8%), PCA (110, 12.0%), RDA (459, 50.2%), and BayeScan (824, 90.2%).

Table 3- Data summary from LFMM, RDA, bayescan and PCAdapt

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Candidate SNPs | SNPs within genes | SNPs not within genes | Matched genes | Non-candidate SNPs | Non-candidates SNPs within genes | Non-candidates SNPs not within genes |
| LFMM | 92 | 73 | 19 | 44 | 17147 | 14863 | 2284 |
| RDA | 822 | 658 | 164 | 459 | 16417 | 14283 | 2134 |
| Bayescan | 1836 | 1501 | 335 | 824 | 15403 | 13440 | 1963 |
| PCAdapt | 173 | 144 | 29 | 110 | 17066 | 14797 | 2269 |
| Total | 2923 | 2376 | 547 | 1437 | 66033 | 57383 | 8650 |

Table 3- contingency table for Chi-square test

|  |  |  |
| --- | --- | --- |
|  | Candidates SNPs | Non-candidates SNPs |
| In Genes | 2376 | 57383 |
| Not in Genes | 547 | 8650 |

The SIFT4G predictions determined that 17 of the 20 overlapping candidate genes contained an environmentally-associated SNP located within the coding region, and 8 of these genes contained nonsynonymous mutations within our dataset (Table 3-13).

Table 3- SIFT4G results for candidate SNPs within Genes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| gene\_ID | total\_SNPs | nonsynonymous | | | synonymous | | |
| total | tolerated | deleterious | total | tolerated | deleterious |
| Ct.00g025080 | 3 | 1 | 1 | 0 | 2 | 2 | 0 |
| Ct.00g026900 | 8 | 0 | 0 | 0 | 8 | 8 | 0 |
| Ct.00g030230 | 2 | 1 | 1 | 0 | 1 | 1 | 0 |
| Ct.00g032480 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Ct.00g049290 | 3 | 2 | 1 | 1 | 1 | 1 | 0 |
| Ct.00g051300 | 3 | 2 | 2 | 0 | 1 | 1 | 0 |
| Ct.00g062900 | 3 | 0 | 0 | 0 | 3 | 3 | 0 |
| Ct.00g064410 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g095350 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Ct.00g154760 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g176220 | 6 | 1 | 1 | 0 | 5 | 5 | 0 |
| Ct.00g179740 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g237940 | 3 | 0 | 0 | 0 | 3 | 3 | 0 |
| Ct.00g238000 | 4 | 1 | 0 | 1 | 3 | 3 | 0 |
| Ct.00g280270 | 5 | 2 | 2 | 0 | 3 | 3 | 0 |
| Ct.00g280280 | 3 | 2 | 2 | 0 | 1 | 1 | 0 |
| Ct.00g290200 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Total | 53 | 12 | 10 | 2 | 41 | 41 | 0 |

Among these genes, several genes are noteworthy due to their intriguing functions. For instance, Ct.00g030230, identified as a seminal plasma protein, may play a role in reproductive success [[132,133]](https://paperpile.com/c/wo71Nx/G1pzW+lyGvQ). Ct.00g095350, a defective proboscis extension response-related (DPR) protein has been shown to affect feeding behavior in fruit flies [[134]](https://paperpile.com/c/wo71Nx/lYeUP). Ct.00g154760, which was annotated as a carnitine O-acyltransferase, is involved in fatty acid metabolism [[135]](https://paperpile.com/c/wo71Nx/tOlnG), a pathway crucial for energy provision under varying climatic conditions. Lastly, Ct.00g04290 encodes a PERIOD CIRCADIAN PROTEIN, which is essential for the regulation of circadian rhythms, aligning life cycle events with environmental cues [[136–138]](https://paperpile.com/c/wo71Nx/TjXw1+E0i1J+Q5Fwo). Notably, within this gene, two nonsynonymous and one synonymous SNPs were discerned. SIFT predictions indicated that one of the nonsynonymous SNPs could be deleterious, potentially impacting the protein's functionality and, by extension, the organism's adaptability to environmental rhythmic changes. This was the one of the two candidate SNPs classified as deleterious by SIFT, so it represents a significant point of interest for its potential role in the ecological adaptation of *Cx. tarsalis*.

## Discussion and Conclusion

We successfully assembled a draft genome for *Cx. tarsalis* and identified hundreds of thousands of polymorphic markers across individuals sampled from diverse locations in North America, primarily in four regions: the West Coast, Southwest, Northwest, and Midwest. Each region is characterized by unique climatic conditions that likely influence genetic variation. The West Coast, known for its warm, seasonally dry climate, contrasts quite significantly with the hot, dry conditions and sparse vegetation of the Southwest. In comparison, the Northwest features a cool, wet climate with lush high vegetation, while the Midwest has a cool, dry climate characterized by a mix of both high and low vegetation. These data revealed a distinct pattern of population structure within this species, with clear differentiation among populations from these regions. Additionally, there was a strong correlation between geographic and genetic distances, highlighting the roles of genetic drift and selection in shaping genetic variation across these varied landscapes.

Our analysis also uncovered a significant link between environmental variables and genetic variation, particularly showing that evaporation rates, and vegetation density, temperature and wind are critical environmental factors with strong associations to genetic differentiation within *Cx. tarsalis* populations. Among these, the identification of 53 SNPs with strong evidence of both selection and environmental correlation across multiple tests suggests that these are most likely to be involved in local adaptation processes. These SNPs, linked to crucial biological functions such as circadian rhythms, reproductive success, feeding habits and fat metabolism, and, lay the groundwork for a detailed exploration of the genetic mechanisms driving adaptation in diverse environmental conditions.

Probably the most interesting SNP uncovered in our study was a deleterious mutation in a gene that we found to be the single copy ortholog of per, a well-studied circadian rhythm gene that encodes the regulatory period protein[[139]](https://paperpile.com/c/wo71Nx/xVKTz). Circadian clock genes are critical for synchronizing physiological and behavioral processes in insects as well as nearly all other living organisms, so a mutation in this gene could be key to the mechanisms that have allowed *Cx. tarsalis* to expand both northward and eastward across the North American continent. Research in other vector species have found that variations in circadian clock genes have a profound impact on mosquito behavior and fitness in both *Cx. pipiens* [[136,137]](https://paperpile.com/c/wo71Nx/TjXw1+E0i1J) and the more distantly related Aedes aegypti [[138]](https://paperpile.com/c/wo71Nx/Q5Fwo), and in *Cx. pipiens* it was found that circadian regulators (including per) were necessary for inducing diapause.

While it has been observed that at least some *Cx. tarsalis* populations must be capable of entering diapause in order to navigate the challenges of seasonal extremes and to traverse significant geographical barriers, such as the Rocky Mountains, the genetic mechanisms underlying this trait have not yet been identified in *Cx. tarsalis*. Our discovery of a candidate allele in the per ortholog represents the first identification of a potential genetic mechanism governing diapause in this species. The elucidation of circadian protein-involved feedback loops, as explored by Shetty et al., further emphasizes the potential universality of these mechanisms across mosquito species. This breakthrough underscores this gene's central role in broader adaptive strategies crucial for ecological adaptation in insect populations.

In addition to the circadian regulator gene, we also identified a handful of other highly significant nonsynonymous mutations in other genes related to traits that are known to be key for insect survival during seasonal changes. For example, Ct.00g030230 was identified as encoding a seminal plasma protein, a family of proteins which are known to influence fertility and post-mating behavior in mosquitoes [[132,133]](https://paperpile.com/c/wo71Nx/G1pzW+lyGvQ). Ct.00g095350 was homologous to the DPR gene in Drosophila, where it has been shown to have a role in regulating feeding behavior [[134]](https://paperpile.com/c/wo71Nx/lYeUP). In fruit flies, the DPR gene is involved in the gustatory (taste) response, particularly in the aversion to salt. In mosquitoes, feeding behavior is critical for both nutrient intake and disease transmission, so this gene may be crucial for not only understanding the ecology of *Cx. tarsalis*, but also its disease transmission dynamics. Finally, we also found a significant mutation in the gene Ct.00g154760, which encodes a carnitine O-acyltransferase. Carnitine O-acyltransferase facilitates the transport of fatty acids into mitochondria for β-oxidation, a process integral to energy production [[135]](https://paperpile.com/c/wo71Nx/tOlnG). Adaptations in Ct.00g154760 that enhance the enzyme's efficiency could provide *Cx. tarsalis* with survival advantages by optimizing energy utilization under fluctuating conditions, and could also be related to diapause behaviors, since insects must accumulate fat reserves prior to entering into the diapause state[[140]](https://paperpile.com/c/wo71Nx/8rXSi).

Overall, our investigation into the *Cx. tarsalis* genome has highlighted the critical importance of environmental adaptations for understanding this mosquito's distribution and adaptation to a wide range of habitats across North America. Identifying genetic markers linked to circadian rhythms, reproductive processes, and metabolic functions reveals how selection on a few key genetic variants may have sculpted the species' ability to adapt to an array of environmental challenges and expand its range both northward and eastward. Such genetic insights are pivotal as climate change continues to reshape habitats, potentially enabling vector species like *Cx. tarsalis* to spread into new areas and present new health challenges in the near future.

# National West Nile Virus Prediction Model

West Nile Virus (WNV) continues to pose significant public health challenges, resulting in considerable human morbidity, mortality, and economic burdens across the United States. Accurate and timely predictive models are critically needed to guide public health interventions, enhance disease surveillance, and optimize vector control strategies, especially given the increasing complexity and variability of transmission patterns due to climate change, urbanization, and ecological shifts.

Chapter 2 outlined the intricate interplay of ecological, climatic, demographic, and genetic factors influencing WNV transmission dynamics, emphasizing the necessity for integrated, multidisciplinary approaches. Chapter 3 further explored these complexities by examining the population genetics of Culex tarsalis—one of the primary vectors of WNV in the United States—highlighting genetic differentiation driven by environmental adaptation and the implications for vector competence and distribution.

Building upon these foundational insights, this chapter explores a national-level predictive model synthesizing historical surveillance data, climatic conditions, vector ecology, and demographic factors. This comprehensive modeling framework seeks to address limitations observed in previous models, which often struggle with accurately predicting outbreaks due to spatial and temporal heterogeneity. By integrating genetic insights from *Culex tarsalis* with extensive epidemiological and environmental data, this chapter investigates the potential and limitations of national-scale predictive modeling, providing insights into broader patterns and informing public health strategies at the national level..

## Data source

### Clinical Disease Surveillance Data

This study utilized data from ArboNET, a national arboviral surveillance system managed by the CDC in collaboration with state health departments[[141]](https://paperpile.com/c/wo71Nx/uArf). ArboNET provides comprehensive surveillance data on arboviral infections, including annual county-level summaries of both human and non-human (birds, mosquitoes, and horses) West Nile Virus (WNV) cases spanning from 1999 to 2023*.*

### Climate Data

Same as in the *Cx. tarsalis* project environmental data section, utilizing ERA5-Land monthly averaged data (1950-present) from the Copernicus Climate Data Store [[106]](https://paperpile.com/c/wo71Nx/fbT8b) to extract annual average and monthly average climate data. Climate data representation for each county is based on county seat coordinates.

The influence of daylight length on West Nile Virus (WNV) dynamics was explored due to its potential role in regulating the biology of *Cx. tarsalis*, the primary vector of WNV. Insights from a population genetics study on *Cx. tarsalis* identified genes associated with circadian rhythm adaptation. Circadian rhythms are closely tied to photoperiods, such as daylight length, and play a critical role in regulating diapause—a physiological state that affects mosquito activity and seasonal disease transmission.

Average monthly daylight length (in hours) was calculated for each record in the dataset using geographic coordinates (latitude and longitude) and temporal data (year and month). These calculations provide an environmental variable to assess its relationship with WNV case dynamics and mosquito behavior.

### Land Cover Data

The land use information is derived from the global 1-km Consensus Land Cover dataset [28], an invaluable resource offering twelve distinct data layers. Each layer represents the prevalence of a specific land-cover type, with values ranging from 0 to 100, denoting the percentage of coverage within a given area. The dataset encompasses a diverse range of land-cover classes, including evergreen and deciduous needleleaf trees, evergreen broadleaf trees, deciduous broadleaf trees, mixed and other tree types, shrubs, herbaceous vegetation, cultivated and managed vegetation, regularly flooded vegetation, urban and built-up areas, snow and ice, barren landscapes, and open water. Spanning latitudes from 90ºN to 56ºS and longitudes from 180ºW to 180ºE, it provides comprehensive global coverage at a resolution of approximately 1 km per pixel at the equator [29], offering a remarkable perspective on land-cover patterns worldwide. Land cover data representation for each county is based on county seat coordinates.

### Birds Data

The eBird dataset, managed by the Cornell Lab of Ornithology, is one of the world’s largest biodiversity-related citizen science initiatives [30]. It aggregates bird observation data contributed by a global community of birders. For this study, eBird data includes observations of common North American bird species, aggregated by species, county, month, and year, along with avian phylodiversity metrics based on 2019 observations. The dataset provides valuable insights into bird distribution, abundance, and trends, with rigorous data quality checks overseen by regional experts[[56]](https://paperpile.com/c/wo71Nx/rS46).

### Demographic Data

Compiled from various sources including the U.S. Census Bureau (population, poverty estimates, and land area data) and Federal Communication Commission (county data). The focus is on data from 2000 to 2021, with manual adjustments for changes such as county name alterations.

## Methodology

### Data Preprocessing

#### Integration of multi-sources data

Harmonizing diverse datasets for machine learning is challenging, especially with the inconsistencies in county identifiers. I use FIPS codes as standard identifiers, yet not all sources include these, and county names often differ between datasets. This requires a thorough manual examination to ensure accurate FIPS code allocation, a task complicated by occasional county renames and FIPS updates. Meticulous effort is essential for successful data integration.

#### Climate data processing

For each data row representing a specific geographic location and year within our study, I employed the xarray library [[142]](https://paperpile.com/c/wo71Nx/5KmDR) in Python to process the ERA5-Land monthly averaged climatic data, which covers the period from 1950 to the present. From this dataset, I extracted the maximum, minimum, and mean values for each climate variable on an annual basis. This approach ensures that I accurately represent the climatic variability and trends for each location over the course of a year, providing a robust foundation for analyzing the influence of climate on West Nile Virus transmission dynamics.

For the monthly report dataset, I extracted the monthly average values for each climate variable within the monthly on the corresponding county. To better understand the past climate variables effect. The average monthly data of the previous one month, two months is also extracted for each location and climate variable.

#### Integration of land cover data

Additionally, I will incorporate land cover variables from the global 1-km Consensus Land Cover dataset [[143]](https://paperpile.com/c/wo71Nx/4QmRI), employing the OpenCV library [[144]](https://paperpile.com/c/wo71Nx/cRPRK) and xarray library [[142]](https://paperpile.com/c/wo71Nx/5KmDR) in Python to extract consensus land cover information pertinent to the county seat coordinates.

### Standardization of data granularity

* Annual Based Dataset

In the context of the data processing for my study, the human West Nile Virus (WNV) data is aggregated on an annual basis at the county level, whereas the non-human WNV data is reported with greater frequency, on a weekly basis, also by county. To harmonize the granularity for comparative and predictive analysis, I standardized all data to an annual county-level scale. This alignment allows for a consistent approach in analyzing the spread and impact of WNV across human and non-human populations, ensuring that the annual cycle of the disease's prevalence is accurately captured, and that the data is comparable across the different categories of subjects affected.

In preparing the dataset for predicting WNV cases, a series of data manipulation and imputation tasks are conducted to ensure data integrity and usability for predictive modeling. Initially, I filtered human infection records from 2004 to 2023 and extracted essential county information, removing duplicates and entries with missing FIPS or State data. Subsequently, I constructed a comprehensive data frame reflecting each unique FIPS code's yearly data, augmented with corresponding state, county, and geolocation details. To address the challenge of missing data, I employed a two-fold imputation strategy: setting case counts to zero in the absence of recorded activity and replacing missing human case reports with the average number of cases over the 13-year span (2004 - 2018) for each county.

### Correlation Analysis using the Mantel Test

Before delving into machine learning, understanding the relationships between environmental variables and the target variable is essential. Kendall’s tau correlation, a robust non-parametric test developed by Maurice Kendall in 1938, will be utilized for this purpose [[145]](https://paperpile.com/c/wo71Nx/SeXz5). Unlike parametric tests, Kendall's tau doesn't assume a specific data distribution, making it ideal for analyzing ecological and environmental datasets.

### Machine Learning / Statistical Models

The study employs various models to predict West Nile virus occurrences, considering the complex interplay between environmental, clinical, and land cover factors. Each model is selected based on its strengths in capturing non-linear relationships, temporal patterns and incorporating spatial dependencies. Criteria for model selection include their ability to handle high-dimensional data, robustness against missing information, and capacity to capture spatial and temporal dynamics intrinsic to disease spread. This selection process, informed by a literature review [[45,56,146–149]](https://paperpile.com/c/wo71Nx/rS46+98UX+Cn7O7+2cTN+lr1FM+Tzj1I), underscores the value of utilizing a combination of traditional machine learning techniques and advanced models to deepen our understanding of disease determinants and improve predictive accuracy.

Table 4-1 summarizes key aspects of each model used in this study, including their descriptions, rationales for selection, as well as pros and cons. Understanding the strengths and limitations of each model is crucial for selecting the most appropriate approach based on the specific requirements and characteristics of the dataset and prediction task.

Table 4- Comparative Overview of Predictive Models for West Nile Virus Prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Description | Rationale for Selection | Pros | Cons |
| Random Forest[[150]](https://paperpile.com/c/wo71Nx/uImM) | Ensemble learning, multiple decision trees | Robustness, handles non-linear data and missing data | Robust to against missing values | Complex, reduced interpretability |
| SVM[[151]](https://paperpile.com/c/wo71Nx/zGuq) | Finds hyperplane separating classes in feature space | High-dimensional effectiveness | Performs well in high-dimensional spaces | Complex parameter tuning |
| Neural Network[[152]](https://paperpile.com/c/wo71Nx/G14k) | Computing system for pattern recognition | Learns complex relationships | Exceptional pattern recognition | Proneness to overfitting |
| Ensemble Model | Combines SVM/RF/HGBR with Autoregression | Leverages both strengths | Integrates spatial and temporal analysis | Constraints on temporal data structure |
| SGLMMs[[153]](https://paperpile.com/c/wo71Nx/HmsX) | Extend GLMs with spatially correlated random effects | Models’ spatial correlation | Accurate spatial dependencies modeling | Complex interpretation |
| HGBR[[154,155]](https://paperpile.com/c/wo71Nx/IuTM+wBdk) | Gradient boosting using histograms for efficiency | Handles large datasets efficiently, captures complex relationship | Fast training, scalable, robust to missing values, interpretable feature importance | Sensitive to hyperparameters, can overfit without tuning |

### Training, Testing and Hyperparameter Tuning

All models, except for the linear model and SGLMMs, were trained on data from 2004 to 2016, hyperparameters tuned on data from 2016 to 2018 (inclusive), and tested on data from 2019 onward. In contrast, the linear model and SGLMMs were trained on data from 2004 to 2018 (inclusive) and tested on data from 2019 onward. This setup ensures that most models undergo hyperparameter optimization on a separate validation set, while the linear model and SGLMMs utilize a longer training period to potentially improve stability and generalizability

The hyperparameter tuning for the models was implemented using the “hyperopt” python package with the Tree-structured Parzen Estimator (TPE) algorithm[[156,157]](https://paperpile.com/c/wo71Nx/KSXC+if03), optimizing for the negative Q² score on a validation set. The best hyperparameters were selected by iterating over 100 evaluations.

For the Support Vector Machine (SVM) model, the tuning focused on the regularization parameter (C) ranging from 0.1 to 10, the epsilon value (0.1 to 5), the kernel type (chosen among *“rbf”*, *“linear”*, *“poly”*, and *“sigmoid”*), and the gamma parameter (either *“scale”* or *“auto”*).

For the Random Forest (RF) model, the tuning involved a broader set of hyperparameters. The number of estimators was optimized between 100 and 1000 (in increments of 10), while the maximum tree depth varied from 1 to 20. The minimum samples required for a split and a leaf were tuned between 2 to 10 and 1 to 10, respectively. The maximum features parameter, which controls the number of features considered for splits, was chosen from both continuous values (0.1 to 1.0) and categorical options (“sqrt”, “log2”). Additionally, max leaf nodes (10 to 100), max samples (0.1 to 1.0), and min impurity decrease (0 to 0.1) were optimized.

For the Histogram-Based Gradient Boosting Regression (HGBR) model, hyperparameter tuning included the maximum depth (1 to 30), number of iterations (100 to 1000, in steps of 100), learning rate (0.01 to 0.5), and L2 regularization strength (0.0 to 1.0). The max leaf nodes (10 to 100, in steps of 10), minimum samples per leaf (1 to 10), and maximum number of bins (10 to 255, in steps of 5) were also optimized. Additionally, the scoring metric was selected from “loss”, “neg\_mean\_squared\_error”, and “neg\_mean\_absolute\_error”.

For the Neural Network (NN) model, hyperparameter tuning included the number of layers (2 to 4), number of neurons per layer (8 to 128, in steps of 8), and dropout rate (0.0 to 0.5). The optimizer used was Adam, with mean squared logarithmic error (MSLE) as the loss function. The best hyperparameters were determined based on validation loss, and the final model selection aimed to balance predictive performance and generalization.

These tuning procedures ensured that each model leveraged the best possible parameter configurations for predicting West Nile Virus (WNV) cases, improving forecasting accuracy and robustness.

### Evaluation Metrics

To comprehensively assess our model's predictive capability for West Nile virus cases, I apply a few evaluation metrics on the testing dataset, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), predictive coefficient of determination (Q²) and bootstrapping technique to calculate the 95% confidence interval. Each metric offers a unique lens through which to view model performance.

Besides the evaluation metrics I mentioned above, I also add a baseline model to compare with each model which is a model that will always predict the result with the mean of the disease count.

## Results

### National CDC Dataset

Figure 4-1 shows the total number of human WNV disease cases annually from the CDC national dataset. The plot shows significant year-to-year variation in WNV neuroinvasive disease cases from 2004 to 2023. The highest peak occurred in 2012, suggesting a major outbreak that year, while smaller peaks are visible in 2006 and 2021. Dips in cases are notable in 2008, 2010, and 2020, which may be influenced by bird immunity cycles and population turnover[[158–160]](https://paperpile.com/c/wo71Nx/09CK+mp9d+33UQ), reducing virus transmission until susceptible bird populations are replenished. The decline in 2020 may also be partially attributed to pandemic-related factors, such as changes in human behavior, healthcare reporting, or vector control efforts. Overall, the data reflects fluctuating case counts without a clear long-term trend.

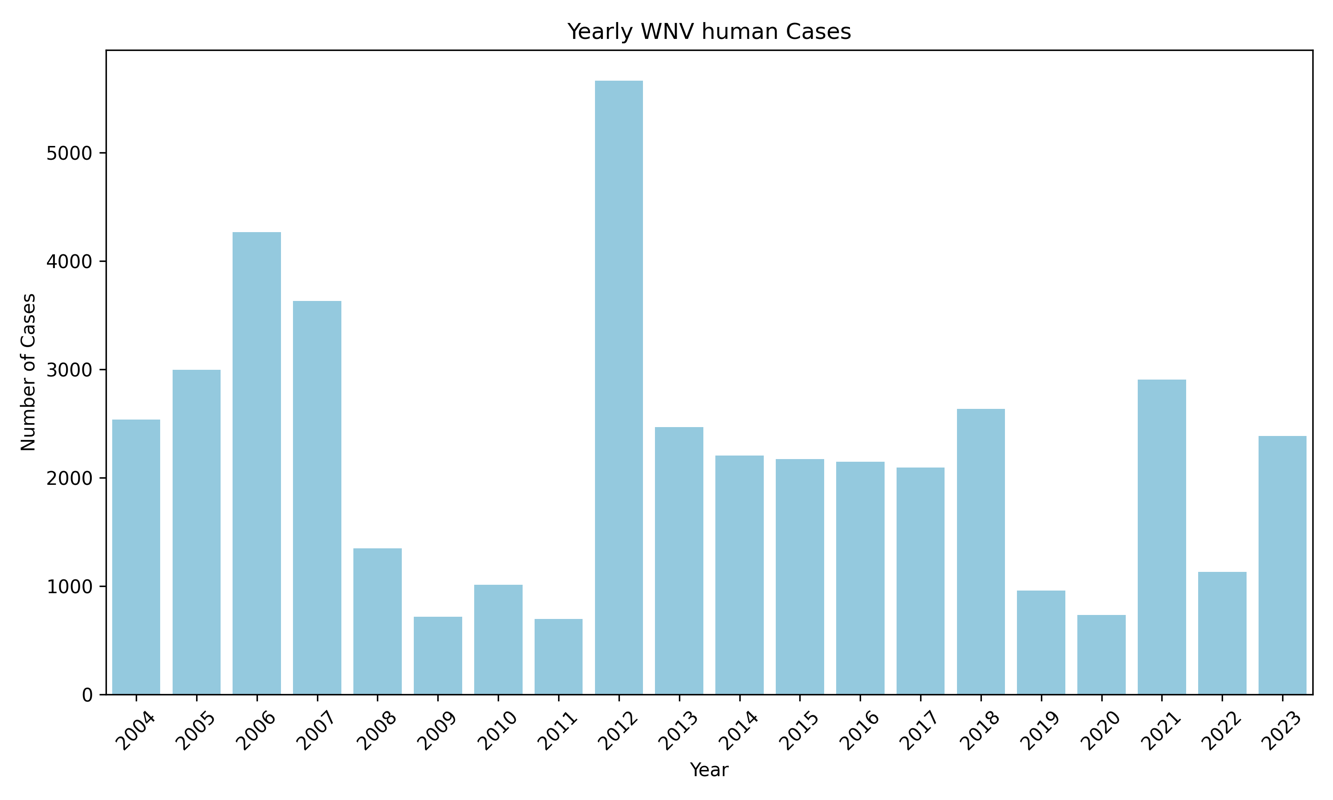


Figure 4- National Annual Total Number of West Nile Virus (WNV) Human Cases (2004–2023)

Figure 4-2 presents Kendall’s tau correlation heatmaps, analyzing the national CDC dataset on annual human West Nile Virus (WNV) cases in relation to environmental and demographic factors. The heatmap indicates that most aggregated environmental variables exhibit minimal correlation with annual WNV case counts. The strongest correlations observed are a negative correlation of -0.17 between deciduous broadleaf tree coverage and WNV human cases and a positive correlation of 0.17 between urban coverage and WNV human cases. This suggests that macro-level environmental metrics may not adequately capture the complex dynamics driving disease transmission.

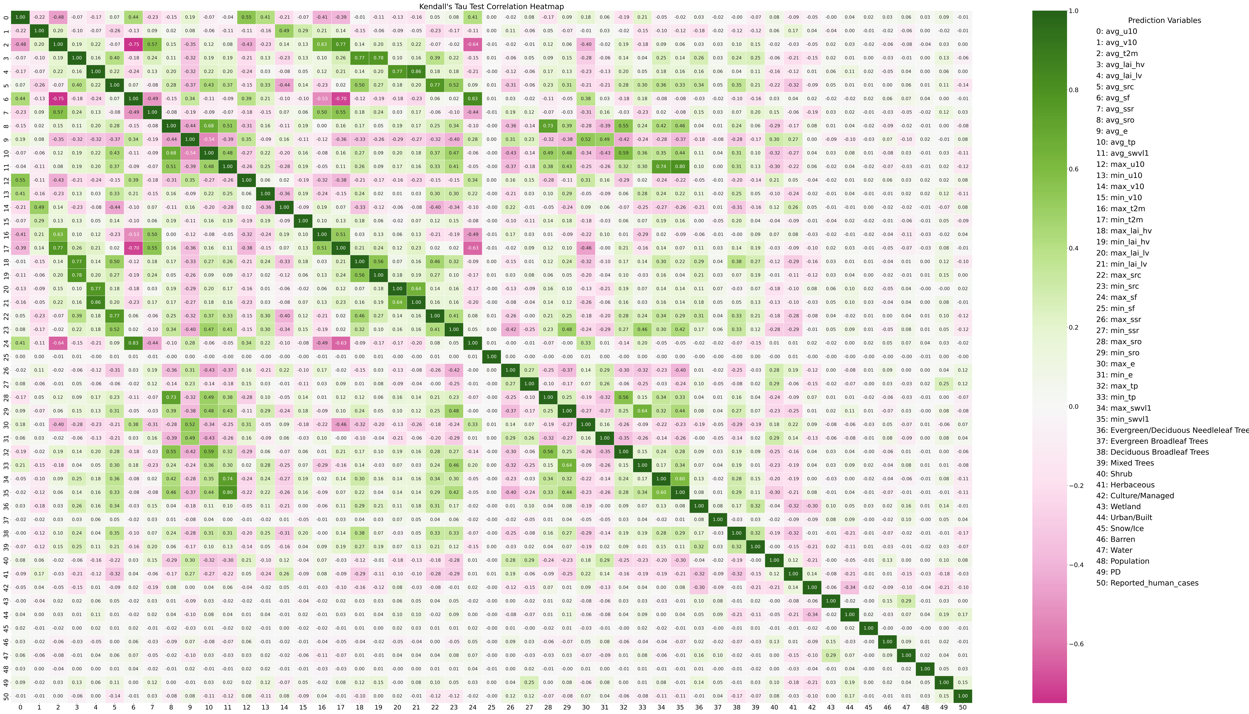
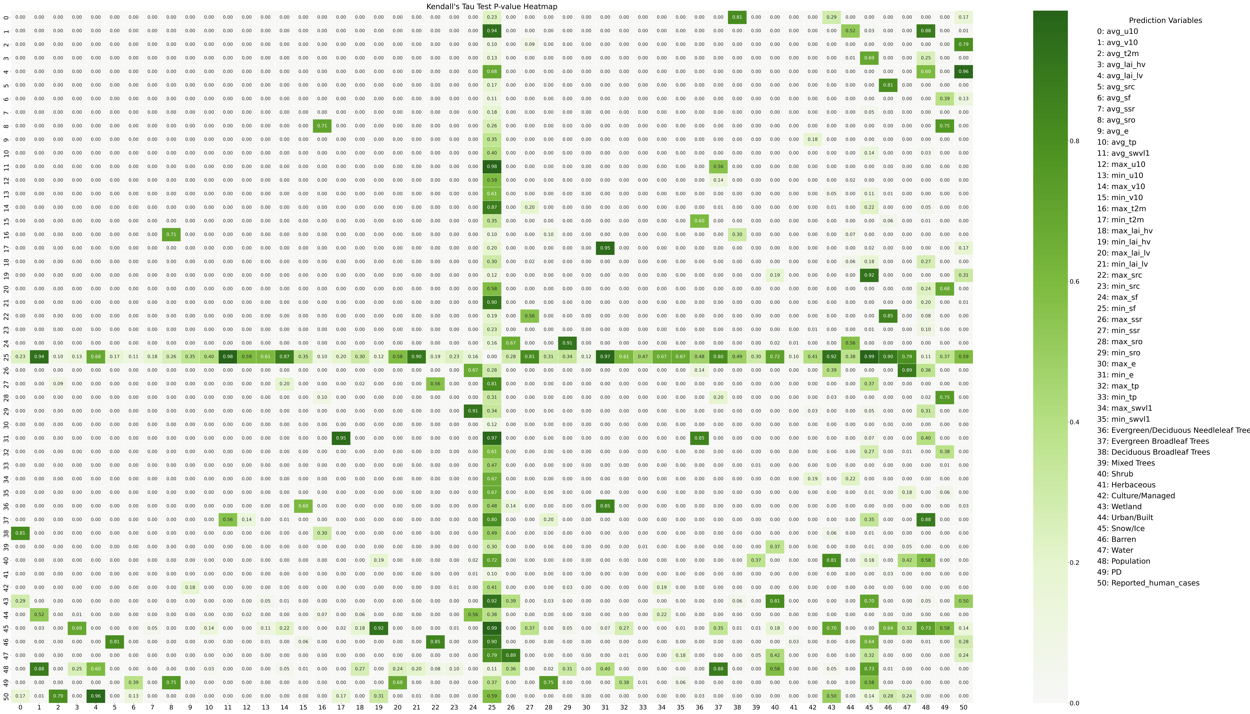


Figure 4- Kendall's Tau correlation Heatmap for National CDC Dataset: Annual Human Cases vs. Environmental Variables (p-values heatmap see Supplemental section)



Following the correlation analysis, an initial modeling approach was undertaken to predict nationwide human WNV cases using various machine learning models. The results, presented in Table 4-2, indicate that predictive performance is generally modest across models, with small variations in accuracy. Among them, the Support Vector Machine (SVM) model demonstrated the lowest Mean Squared Error (MSE) of 192 and the highest Q² score of 0.09, suggesting it offers a slight predictive advantage over other approaches. However, the low Q² scores across all models indicate that environmental variables alone may not be sufficient for accurately forecasting WNV human cases.

Table 4- Performance of Different Predict Models on National Model

|  |  |  |
| --- | --- | --- |
| Model | MSE | Q2 |
| Linear Regression | 199 | 0.06 |
| Random Forest \* | 202.11 | 0.04 |
| SVM | 192 | 0.09 |
| Neural Network | 202.10 | 0.04 |
| SGLMM \*\* | 210.89 | 0.01 |
| HGBR | 198 | 0.06 |

\* The MSE and Q2 for RF is average results of 100 runs

\*\* Simplify the variables only use Water retention capacity + Surface net solar radiation + Urban Built + Avian Phylodiversity

While tuning the prediction models for West Nile Virus (WNV) incidence, several significant data-related challenges have emerged:

**Inconsistency in Data Collection**: The collection of WNV case reports varies considerably across states, presenting substantial hurdles for modeling efforts at a national level. This inconsistency can lead to skewed data representations, which may compromise the accuracy and reliability of predictive analytics.

**Significant Class Imbalance**: The imputation of missing values exacerbates the existing imbalance between reports of zero cases and those documenting one or more instances of WNV. Such disproportionality poses a risk of biasing the model towards predicting the majority class (0 disease case) and necessitates the adoption of specialized techniques to achieve a balanced representation of case occurrences.

**Data Granularity**: An additional complexity encountered in predicting annual WNV cases is the granularity of environmental data. When forecasting yearly disease counts, environmental variables must be aggregated—typically as annual averages, maximums, or minimums. This approach, while necessary for aligning temporal scales, may obscure finer temporal correlations between environmental conditions and WNV transmission patterns. For instance, the use of yearly averages or extremes of temperature does not capture the subtleties of seasonal fluctuations that could be critical in understanding and predicting WNV outbreaks. The nuances of how and when these environmental factors influence mosquito populations and virus transmission may be lost, potentially diminishing the predictive power of our models.

Each of these issues requires tailored methodological adjustments to ensure the robustness of the predictive models. The following sections detail the specific approaches undertaken to mitigate these challenges and discuss the results yielded by these refined modeling strategies.

### Challenges in National Model

* Inconsistency in Data Collection

Given the variability in data collection and reporting of West Nile Virus (WNV) cases among states, we adopted a state-centric modeling approach. Instead of building a model for the entire country, we build a model for each individual state. Figure 4-3 presents the outcomes of employing Support Vector Machine (SVM) algorithms to predict West Nile Neuroinvasive Disease (WNND) occurrences for each state. The displayed models, limited to those achieving a positive Q2 score, indicate a disparity in performance across states. For instance, California, Illinois, Michigan, and New York demonstrate superior model efficacy compared to those of Pennsylvania, North Dakota, and Washington, suggesting that the robustness of local WNV surveillance systems may significantly influence predictive accuracy.

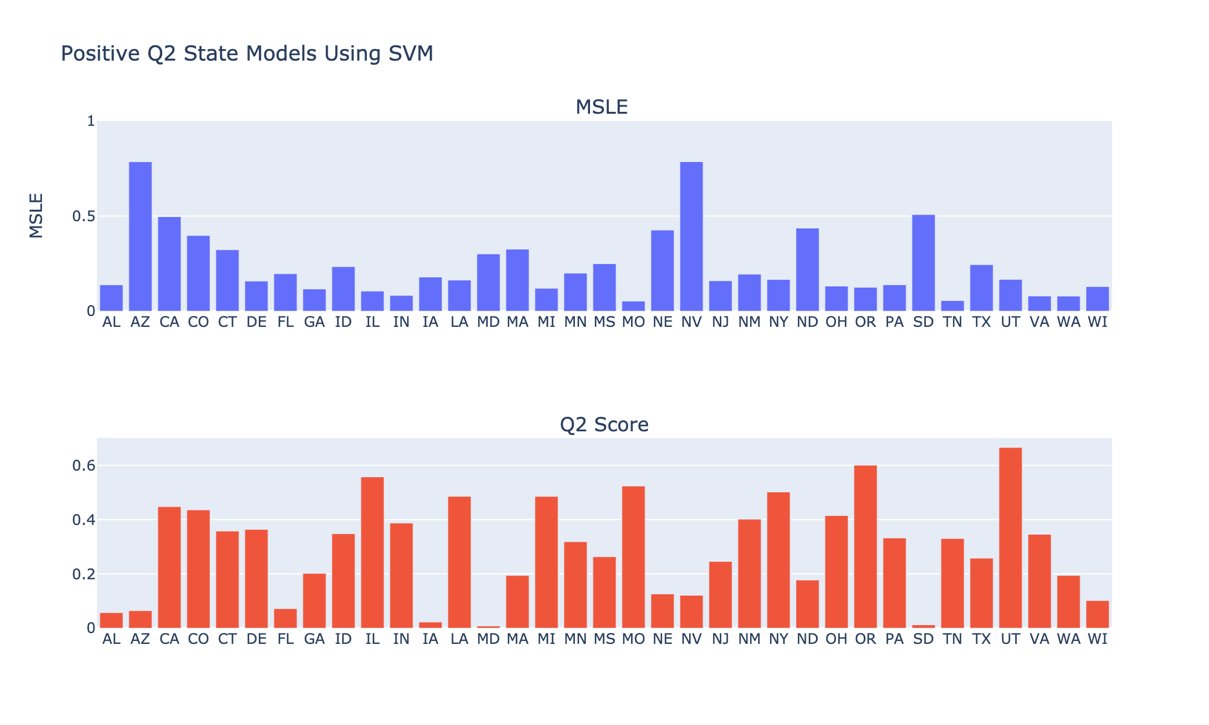


Figure 4- Comparative Analysis of SVM Models for WNV cases Across States

As part of an integrated strategy, I shifted my analytical focus from the quantitative assessment of West Nile Virus (WNV) incidences to a binary determination of WNV presence within individual counties. This pivot required reconceptualizing the issue from a regression framework to a classification paradigm. The newly adopted classification approach was applied to data from New York, California, and Illinois, which were specifically chosen due to their robust WNV case reporting systems that include both human and non-human instances. The efficacy of the SVM-based classification models is detailed in Table 4-3, where they are evaluated against a national model. These comparisons shed light on the models' capacity to accurately predict binary outcomes indicative of WNV presence.

To enhance model performance, I explored a range of probability thresholds from 0 to 1 to distinguish between zero cases (0) and non-zero cases (non-0). The optimal threshold was selected based on its ability to maximize the F1 score, a harmonic mean of precision and recall that balances the trade-off between false positives and false negatives. Among the state models, California's demonstrated the most robust performance. Conversely, the national model's high accuracy, coupled with a lower F1 score, highlights the persistent challenge posed by class imbalance within the dataset.

Table 4- SVM Predictive Outcomes for WNV Surveillance in State models,

New York, California, Illinois, Colorado and National Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Threshold | F1 | Accuracy | AUC-ROC |
| New York | 0.58 | 0.694 | 0.846 | 0.82 |
| California | 0.85 | 0.860 | 0.826 | 0.87 |
| Illinois | 0 | 0.552 | 0.381 | 0.60 |
| Colorado | 0.33 | 0.743 | 0.842 | 0.85 |
| National | 0.33 | 0.499 | 0.764 | 0.74 |

* Significant Class Imbalance

Figure 4-4 displays the distribution of West Nile Virus Neuroinvasive Disease (WNND) counts nationally. The data reveal a predominant number of zero-case instances compared to non-zero cases. Such an imbalance skews prediction models towards forecasting zero cases, resulting in overly conservative estimates. To counteract this, a subsampling technique was applied to down sample the majority class (zero cases) to align with the minority class (non-zero cases), aiming for a balanced dataset that would enable more accurate predictions.

A graph of a number of numbers

Description automatically generated

Figure 4- WNV case distribution across the nation

To address this, a subsampling technique was applied, downsampling the majority class (zero cases) to achieve a balanced dataset for national-level predictions. However, as shown in Table 4-4, this adjustment did not significantly improve predictive performance—Q² scores remained consistent, while Mean Squared Error (MSE) increased, indicating that models became less conservative but did not substantially enhance accuracy.

Given the limited success of the national model with downsampling, an alternative approach was explored: training individual state-specific models to predict annual WNV case counts at the state level. Figure 4-5 presents the performance of state-specific SVM models, revealing that California, Colorado, Illinois, Michigan, New York, and Utah achieved relatively strong predictive power (Q² > 0.4).

Figure 4-6 further illustrates the annual case change across states, revealing that California, Colorado, Illinois, Louisiana, Michigan, New York and Utah have had a more consistent surveillance reporting over the years where other states have less WNV case reports. This suggests that states with a greater prevalence of WNV cases may provide more robust training data, leading to improved model performance. These findings indicate that WNV transmission dynamics are better captured at a localized scale, where state-specific environmental and epidemiological patterns contribute to more accurate predictions.

Table 4- Performance of Different Predict Models on National Model Using Subsampling

|  |  |  |
| --- | --- | --- |
| Model | MSE | Q2 |
| Linear Regression | 200.46 | 0.05 |
| Random Forest 1 | 202.52 | 0.05 |
| HGBR | 195.42 | 0.07 |
| SVM | 193 | 0.089 |

1 The MSE and Q2 for RF is average results of 100 runs

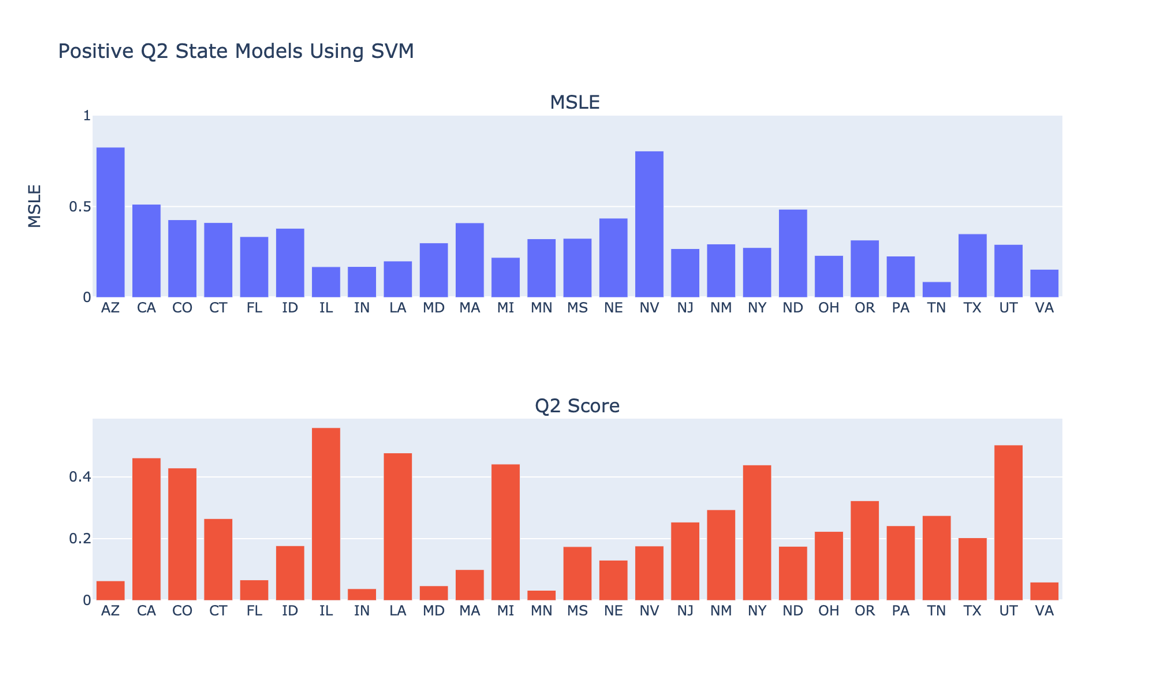


Figure 4- Comparative Analysis of SVM Models for WNV cases Across States Using Subsampling

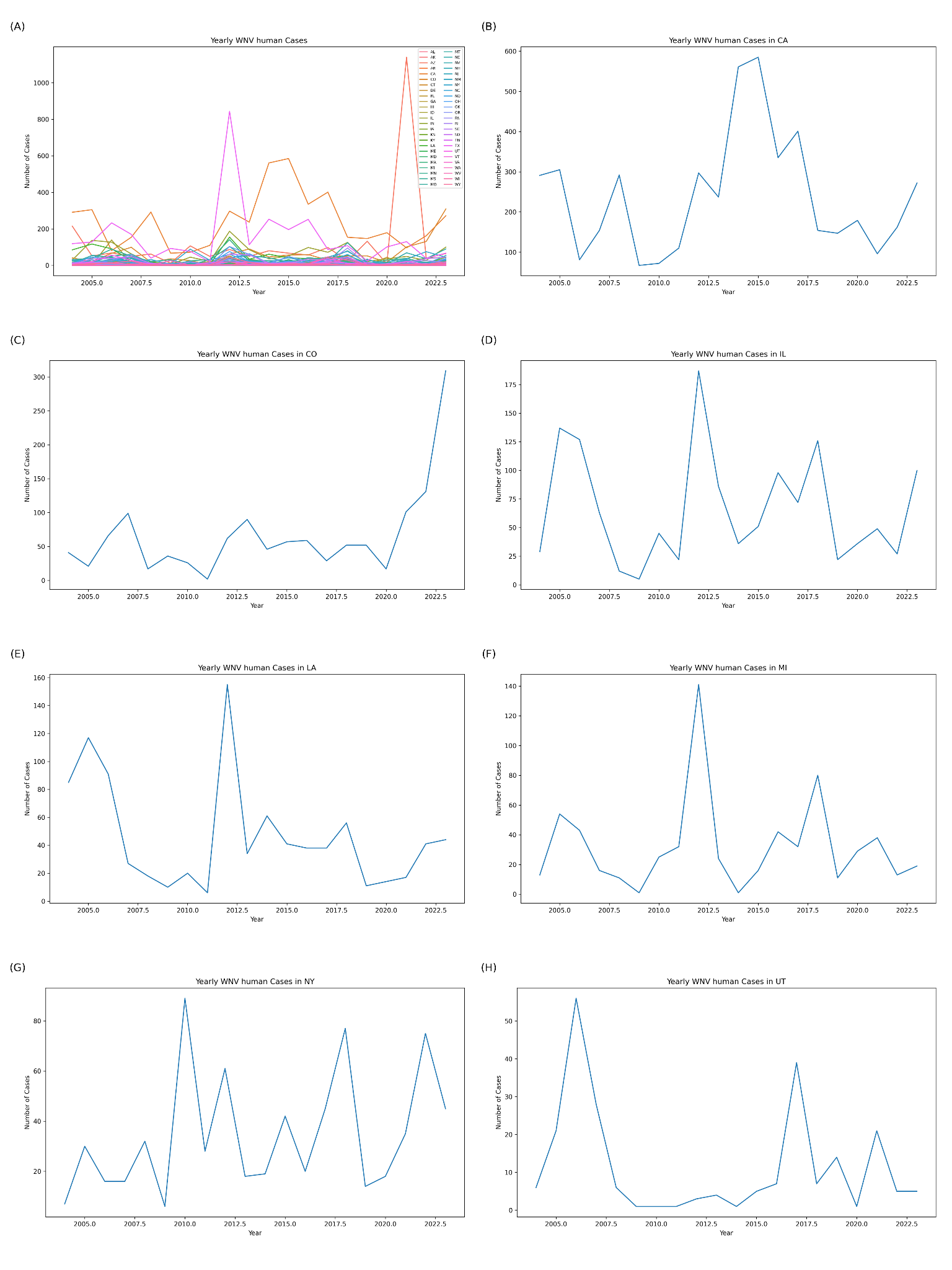


Figure 4- Annual human West Nile Virus (WNV) case counts (2004–2023) for selected U.S. states. (A) Annual case counts for individual states; (B) - (H) Annual case counts for California, Colorado, Illinois, Louisiana, Michigan, New York, Utah

# Regional West Nile Virus Prediction in California

To investigate how data granularity impacts model accuracy, we transitioned from using national CDC ArboNET data, which reports yearly WNV case counts at the state level, to more detailed state-specific datasets. The limitations of the CDC data, primarily its low spatial and temporal resolution, resulted in poor forecasting performance.

## Data source

### Surveillance data

To address these challenges, we utilized two independent datasets with finer granularity, both focusing on California. The first dataset, referred to as the CDPH-Official dataset, was obtained through an official request to the CDPH Vector-Borne Disease Section and includes weekly WNV disease reports for 13 counties (see Clinical Disease Surveillance Data section) between 2004 and 2023. The second dataset, referred to as the CDPH-Scraped dataset, was independently compiled by scraping more than 760 PDF files of weekly reports available on the CDPH website ([westnile.ca.gov/resources\_reports?report\_category\_id=6](https://westnile.ca.gov/resources_reports?report_category_id=6)) . This dataset includes all counties in California and provides a comprehensive view of WNV case distribution across the state. These two datasets, treated independently, offered complementary insights into the temporal and spatial patterns of WNV cases at the weekly and county levels.

* Official CDPH Data:

This data was officially requested from the CDPH Vector-Borne Disease Section. It includes records specifically for 13 counties: Fresno, Kern, Los Angeles, Merced, Orange, Placer, Riverside, Sacramento, San Bernardino, San Joaquin, Solano, Stanislaus, and Tulare. Figure 5-1 shows the annual geographic distribution of human WNV cases in California from 2004 to 2023.

A screenshot of a map

AI-generated content may be incorrect.

Figure 5- Annual Geographic Distribution of Human West Nile Virus Cases in California (2004–2023)

* Scraped Data from CDPH Website:

This dataset was independently compiled by scraping more than 760 PDF files of weekly reports available on the CDPH website (<https://westnile.ca.gov/resources_reports?report_category_id=6>). It contains data for all counties in California and is entirely separate from the official CDPH data.

Figure 5-2 shows the distribution of absolute differences in human WNV disease case counts between two different sources: the official CDPH dataset and the independently scraped dataset. The x-axis represents the magnitude of the case differences, while the y-axis indicates the frequency of these differences. This visualization helps identify the extent and frequency of discrepancies between the two data sources across counties, years, and months. The histogram shows that most case differences between the two datasets (CDPH and indecently scraped dataset) are close to zero, indicating that the datasets are largely consistent. However, a small number of large differences are evident, particularly in counties and months critical for WNV transmission. Table 5-1 shows the top 20 largest case differences, highlighting the specific counties, years, and months where these differences occur. The result shows that Los Angeles County dominates the top differences, with multiple instances of significant differences, especially during late summer and fall months (August to October) in years like 2014, 2015, and 2017. Notably, the largest single difference is 97 cases in Los Angeles County in October 2017, followed by 94 cases in Orange County in August 2014. These outliers suggest potential differences in data collection or reporting methodologies, particularly during peak transmission periods. While the datasets are largely aligned, addressing these discrepancies through further investigation could enhance data reliability and ensure more accurate modeling and surveillance of WNV trends.

A graph with numbers and a bar

AI-generated content may be incorrect.

Figure 5- Histogram plot of WNV cases different between CDPH and independently scraped data

Table 5- Top 20 WNV case differences between CDPH and independently scraped data

|  |  |  |  |
| --- | --- | --- | --- |
| County | Year | Month | Case Absolute Difference |
| Los Angeles | 2017 | 10 | 97 |
| Orange | 2014 | 8 | 94 |
| Los Angeles | 2015 | 9 | 89 |
| Los Angeles | 2014 | 10 | 85 |
| Los Angeles | 2014 | 9 | 77 |
| Orange | 2014 | 10 | 76 |
| Los Angeles | 2014 | 8 | 63 |
| Los Angeles | 2017 | 8 | 60 |
| Los Angeles | 2004 | 8 | 58 |
| Riverside | 2015 | 10 | 58 |
| Los Angeles | 2015 | 11 | 57 |
| Riverside | 2015 | 8 | 48 |
| Los Angeles | 2017 | 9 | 47 |
| Los Angeles | 2023 | 8 | 46 |
| Los Angeles | 2015 | 12 | 45 |
| Los Angeles | 2016 | 8 | 45 |
| Los Angeles | 2014 | 11 | 44 |
| San Bernardino | 2004 | 7 | 42 |
| Los Angeles | 2013 | 10 | 40 |
| Stanislaus | 2023 | 8 | 37 |

### Mosquitoes’ abundance data

The mosquito abundance data, covering all species from 2004 to 2023, was obtained from the California Vectorborne Disease Surveillance System (CalSurv). It includes monthly data for 13 counties (Fresno, Kern, Los Angeles, Merced, Orange, Placer, Riverside, Sacramento, San Bernardino, San Joaquin, Solano, Stanislaus, and Tulare). The data was collected using two trap types: carbon dioxide-baited traps and gravid mosquito traps.

Access to this data is governed by the California Vectorborne Disease Surveillance Data Policy. Researchers or individuals interested in using this data must adhere to the policy outlined in the official document, which can be accessed at: <https://vectorsurv.org/assets/files/calsurv_data_policy.pdf>.

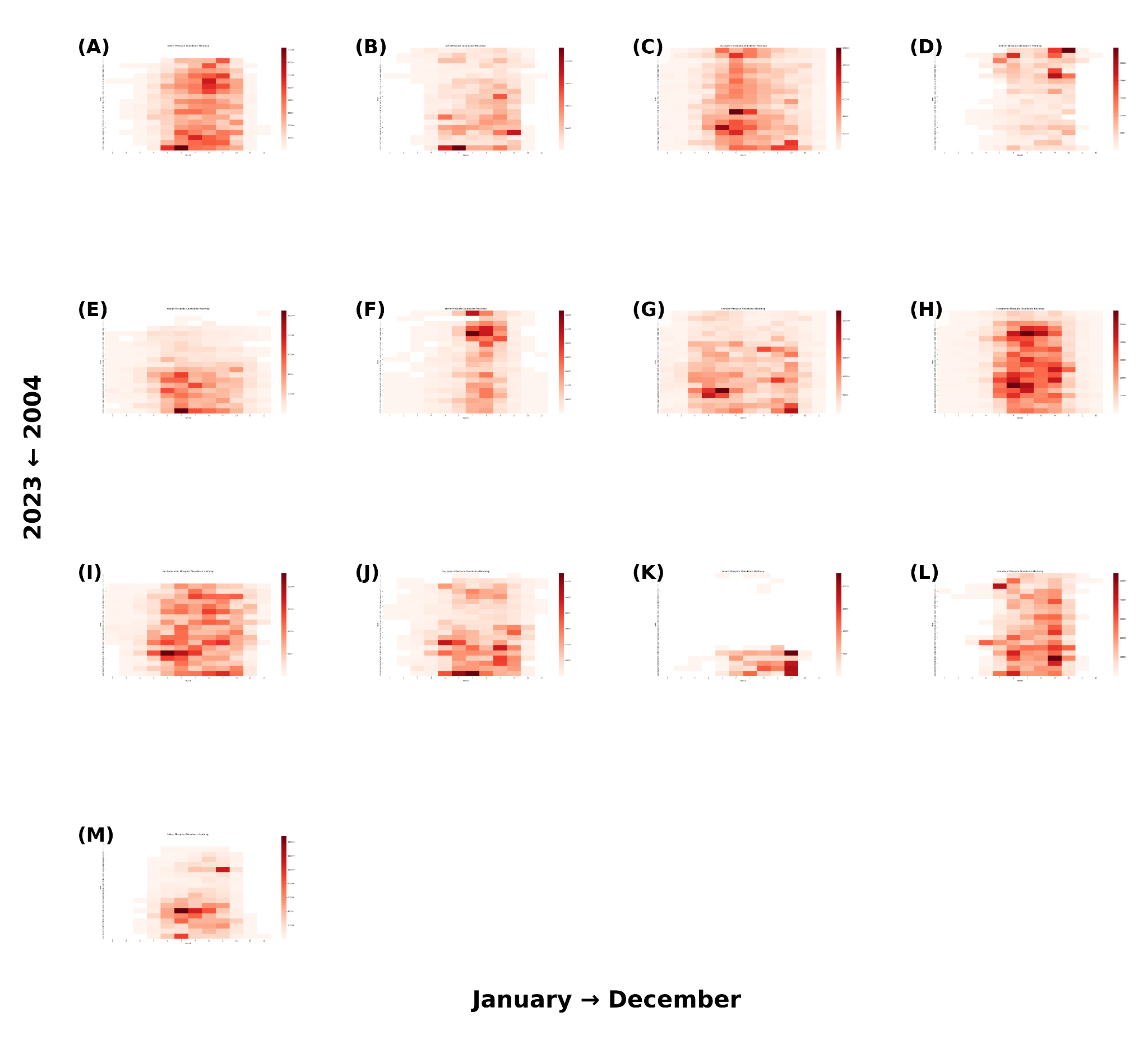
The data display clear gaps in mosquito collections during the early years in some counties, which could make using this data for prediction challenging (See in figure 5-3). These gaps may introduce biases or reduce the reliability of predictive models, particularly when attempting to analyze temporal trends or make comparisons across different counties. Careful handling of missing data, such as imputation or the exclusion of affected time periods, is necessary to ensure robust and accurate forecasting results.

Figure 5- Heatmap of Monthly Mosquito Abundance (2004–2023) Across 13 California Counties, by Month. (A) Fresno; (B) Kern; (C) Los angeles; (D) Merced; (E) Orange; (F) Placer; (G) Riverside; (H) Sacramento; (I) San bernardino, (J) San joaquin; (K) Solano; (L) Stanislaus; (M) Tulare

### Data preprocessing

The data integration and environmental data preprocessing is the same as mentioned in Chapter 4. For the monthly human case reports in California, the dataset was standardized in a manner similar to the annual dataset. A comprehensive data frame was constructed to reflect each unique Federal Information Processing Standards (FIPS) code and month in California from 2004 to 2023.

To address missing data, I consulted the epidemiologists at the California Department of Public Health (CDPH), who confirmed that missing records should be treated as zero cases. Accordingly, all missing entries were filled with zeroes to accurately represent the absence of reported cases, rather than incomplete data collection. It is important to note that the case counts reported here likely underestimate the true number of infections due to asymptomatic and unreported cases.

This standardization process provides a consistent temporal and spatial framework for analysis, minimizing bias from gaps in reporting and enhancing the reliability of subsequent modeling efforts.

### Evaluation metrics

To comprehensively assess our model's predictive capability for West Nile virus cases, I apply a few evaluation metrics on the testing dataset, including Mean Squared Error (MSE), Mean Absolute Error (MAE), predictive coefficient of determination (Q²) and bootstrapping technique to calculate the 95% confidence interval. Each metric offers a unique lens through which to view model performance.

In addition to evaluating the importance of variables and understanding their contributions to the predictive performance of the machine learning models, we used SHAP (SHapley Additive exPlanations) values. SHAP is a widely adopted approach for interpreting complex models by quantifying the impact of each feature on model predictions. Derived from cooperative game theory, SHAP values provide an equitable and consistent measure of feature importance.

SHAP values allow for both local and global interpretability. Locally, they explain individual predictions by attributing the output to input features, making it clear which variables contribute positively or negatively. Globally, SHAP values are aggregated across the dataset to highlight the overall importance of each variable, offering insights into which features most strongly influence the model’s behavior.

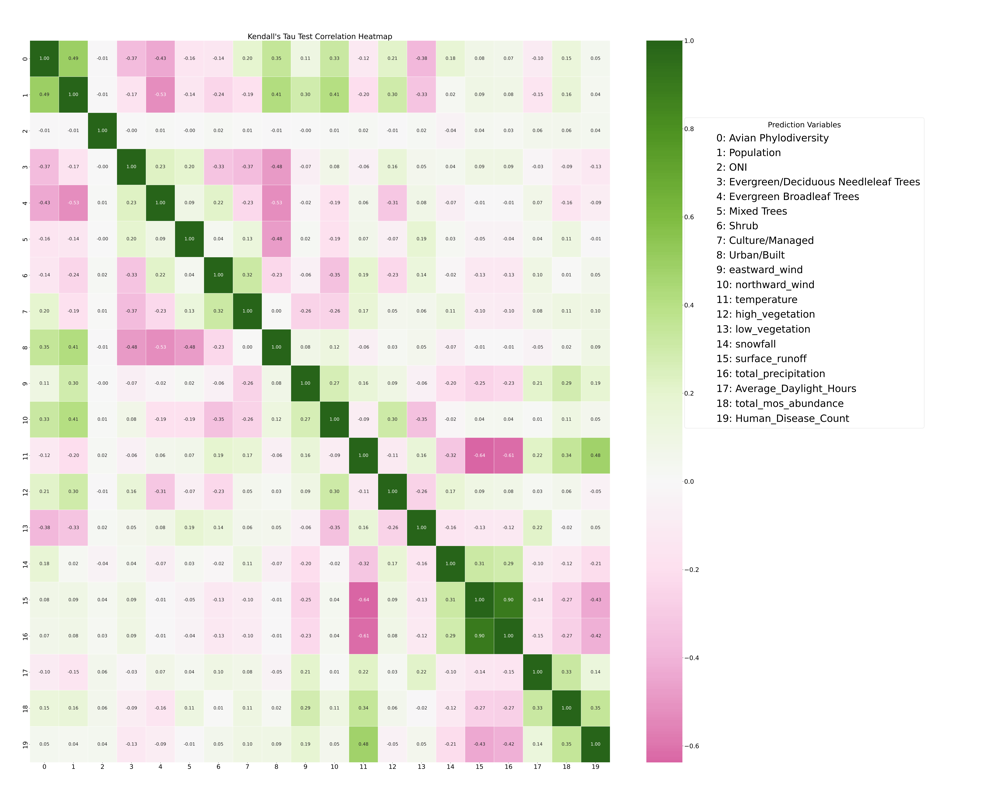
The application of SHAP values in this study has several benefits:

* Enhanced Interpretability: By attributing predictions to specific features, SHAP values clarify the model’s decision-making process.
* Feature Selection: Variables with low SHAP importance can be identified and excluded, potentially improving model efficiency and reducing overfitting.
* Interaction Insights: SHAP values reveal interactions between features by showing how the presence of one variable affects the contribution of another.

In this study, SHAP values were computed for all models to ensure an in-depth understanding of variable importance. This not only supports model evaluation but also aligns the results with biological and ecological domain knowledge, ensuring that the models are both predictive and interpretable.

## Results

### Exploratory data analysis

The enhanced granularity of these datasets revealed stronger and more noticeable connections between WNV case counts and predictors—including environmental variables and demographic factors—at the county level. Figure 5-4 illustrates the correlation improvements observed when using CDPH-Official data compared to national-level CDC data. Figure 5-4 reveals significant positive relationships between temperature, total mosquito abundance, eastward wind, and human disease count, as well as significant negative relationships between surface runoff, total precipitation, snowfall, and human disease count. These findings suggest that these variables are most likely good predictors for WNV incidence. Additionally, average daylight hours and cultural and urban areas exhibit moderate positive correlations, while evergreen tree cover (variables 3 and 4) shows moderate negative correlations, further emphasizing their importance in modeling disease trends. The p-value heatmap confirms the statistical significance of these relationships. These insights highlight the importance of matching data granularity to the scale of environmental and epidemiological processes to enable more accurate forecasting.

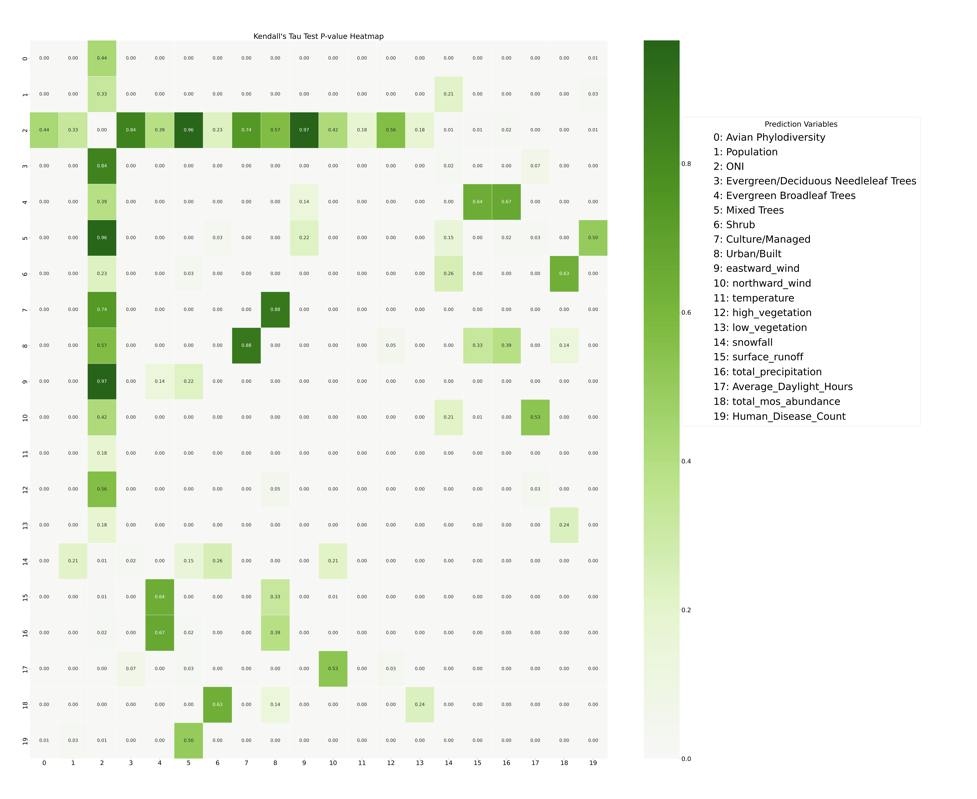


Figure 5- Kendall’s Correlation and P-value Heatmap between environmental variables and WNV disease count using CDPH-Official dataset

Rolling statistics were calculated for West Nile Virus (WNV) human disease counts using a 3-month window to capture short-term trends and variability in case numbers. This approach involved computing the rolling sum, mean, and standard deviation of cases over consecutive 3-month periods, providing insights into temporal patterns, such as seasonal fluctuations and periods of increased or decreased case activity. These rolling statistics help smooth out noise in the data while highlighting meaningful changes in case trends over time. The results visualized in the Figure 5-5, reveal seasonal peaks and variations in WNV cases over the 2004–2023 period, providing a clearer understanding of the temporal dynamics of WNV outbreaks. This approach highlights both the magnitude of outbreaks and the consistency of their timing, helping to visually see the patterns in disease occurrence.

A graph of statistics with numbers

AI-generated content may be incorrect.

Figure 5- Rolling statistics for number of WNV cases in California

### Iterative prediction framework

We implemented an iterative prediction framework to ensure that our models continuously adapted to new data while maintaining a realistic forecasting setup. The process began by training the model on data from the first available year (2004) and optimizing its hyperparameters using data from the following year (2005) through grid search. This optimization step involved systematically evaluating multiple hyperparameter combinations to identify the best-performing configuration based on a predefined evaluation metric (see Section 4.2.4 Hyperparameters tuning section). Once the optimal hyperparameters were determined, the tuned model was used to predict each subsequent year individually from 2006 to 2023. The process was then repeated iteratively: after training on data from 2004–2005, hyperparameter tuning was conducted on data in 2006 using grid search, and the model was applied to predict 2007–2023. This iterative procedure continued, incrementally incorporating all prior years for training, tuning hyperparameters on the most recent available year, and using the optimized model to forecast future cases.

By progressively refining the model with the latest data while leveraging historical trends, this iterative approach allowed us to adapt to evolving patterns in West Nile Virus (WNV) outbreaks, ultimately enhancing the robustness and accuracy of our predictions. Figure 5-6 presents the Q² results for Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR). For improved visualization, Q² values lower than -1 were capped at -1. In Figure 5-6, the y-axis represents the year used for hyperparameter tuning, while the x-axis denotes the predicted year.

The results in Figure 5-6 indicate that predictions for the years 2009, 2010, 2018, 2019, and 2021 consistently exhibit poor performance. While the low predictive accuracy in earlier years (e.g., 2009 and 2010) may be attributed to the limited training data available, which hinders the model’s ability to generalize patterns effectively, this explanation does not account for the poor performance observed in 2018, 2019, and 2021. These findings suggest that additional factors, not captured by the model, may have specifically influenced these years, warranting further investigation.A screenshot of a graph

AI-generated content may be incorrect.

Figure 5- Iterative Forecasting Performance (Q2 Values) of SVM, RF, and HGBR Models

### Fixed training approach

As an alternative approach, we partitioned the dataset into three subsets: 80% of the data from 2004 to 2018 for training, the remaining 20% as a validation set, and 2019 to 2023 as the test set. The models were trained on the training set, hyperparameter tuning was performed on the validation set, and final evaluations were conducted on the test set.

Figure 5-7 presents the Q² and Mean Squared Error (MSE) results for Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR). Among the models, SVM demonstrated the best performance; however, its Q² value was only 0.2, highlighting the model’s limitations in capturing the underlying patterns effectively.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 5- Performance of SVM, RF, and HGBR When Trained on 2004–2018 and Tested on 2019–2023

The poorer performance of the model trained on 2004–2018 and tested on 2019–2023 compared to the iterative approach suggests the presence of concept drift and temporal changes in the data. Factors such as climate variations, shifts in mosquito and bird populations, and changes in human intervention strategies may have altered WNV transmission patterns after 2018, making older data less predictive of recent years. Additionally, the iterative approach adapts yearly, leveraging the most recent data for hyperparameter tuning, whereas the fixed 2004–2018 model relies solely on historical patterns that may no longer be relevant. Overfitting to older trends, data distribution shifts, and the presence of anomalous years (e.g., 2018, 2019, and 2021) further contribute to the model’s reduced generalization ability. These findings indicate that recent data is more informative for forecasting than older historical data, highlighting the need for models that can dynamically adjust to evolving conditions.

### Hybrid autoregressive approach

To improve the accuracy of West Nile Virus (WNV) predictions, we incorporated time-series data into our modeling approach (Figure 5-8). We first calculated the month-to-month difference in WNV cases and created a new column to capture these changes. Next, we generated lagged features by shifting this difference by 1, 2, and 3 months, resulting in columns representing t-1, t-2, and t-3. These lagged variables were then integrated with the environmental factors used in our previous model, forming an enriched dataset for training and testing.

To leverage both historical case trends and external environmental factors, we developed a hybrid autoregressive model that integrates classical machine learning methods with time-series autoregression techniques. By incorporating temporal dependencies, this approach allows the model to better capture patterns in WNV outbreaks, ultimately enhancing prediction accuracy.

A screenshot of a table

AI-generated content may be incorrect.

Figure 5- Example of construction of Lagged Temporal Features for WNV Prediction

To evaluate the effectiveness of our predictive models, we applied Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR) using a structured training, validation, and testing approach. Specifically, we trained the models on data from 2004 to 2015, allowing them to learn historical patterns and relationships. Next, we performed hyperparameter tuning using data from 2016 to 2018, optimizing model performance through grid search to identify the best parameter configurations. Finally, we assessed the models’ predictive accuracy by testing them on an unseen dataset from 2019 to 2023. The results of this evaluation are presented in Figure 5-9. Additionally, we plotted the global SHAP values for each predictive variable across all models (Figure 5-10) to interpret feature importance.

From Figure 5-9, we observe that incorporating autoregression into the model significantly enhances prediction accuracy compared to models that rely solely on environmental variables. Among the models tested, Random Forest demonstrates the strongest performance, achieving a Q² close to 0.5 and an MSE below 13, indicating a more reliable representation of observed case trends. This suggests that integrating temporal dependencies, such as past case trends, provides valuable predictive power beyond environmental factors alone.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 5- Result of Autoregression model using SVM, RF and HGBR

Figure 5-10 further highlights the most influential predictors across the three models, revealing that t-1 (the difference in cases from the previous two months), temperature, average time length, and northward wind play the most significant roles in forecasting case numbers. Notably, temperature, average time length, and northward wind are not only key predictors in our models but also fundamental factors influencing the environmental adaptation of *Culex tarsalis*, as discussed in the previous chapter. Their strong predictive importance in this study reinforces their role in shaping mosquito population dynamics and, consequently, the transmission risk of West Nile Virus.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 5- Feature Importance Analysis Using Global SHAP Values for SVM, RF, and HGBR Models

To determine the optimal number of lagging months to incorporate into the model, we evaluated the Random Forest model using lag values ranging from 0 months (no lag) to 5 months. Figure 5-11 presents the model’s performance across different lag configurations. The results indicate that incorporating t-1 (1-month lag), t-2 (2-month lag), and t-3 (3-month lag) alongside environmental variables yields the highest Q² and the lowest MSE, suggesting that including short-term historical trends significantly enhances predictive accuracy. Beyond t-3, additional lagging months do not provide further improvements, likely due to diminishing temporal dependencies in WNV case trends.

A graph with lines and numbers

AI-generated content may be incorrect.

Figure 5- Impact of Temporal Lag on Random Forest Prediction Accuracy (Q² and MSE)

To quantify the uncertainty in our predictions, we conducted bootstrapping on the SVM, RF, and HGBR models. This involved resampling the test dataset (2019 - 2023) 1000 times with replacement and recalculating the model estimates, allowing us to derive a 95% confidence interval for the predictions. The results are shown in Figure 5-12. This statistical measure provides insight into the variability and reliability of our forecasts, ensuring robustness in our findings.

A group of blue and red graphs

AI-generated content may be incorrect.

Figure 5- Histogram of 95% Confidence Intervals of Q² from Bootstrapped Predictions for SVM, RF, and HGBR. SVM: (A) and (B); Random Forest: (C) and (D); Histogram-Based Gradient Boosting Regression: (E) and (F)

## Discussion

This study demonstrates the effectiveness of incorporating autoregressive components into machine learning models for West Nile Virus (WNV) prediction, significantly improving forecast accuracy over models relying solely on environmental variables. By integrating past case trends as predictive features, our hybrid autoregression approach enhances the temporal resolution of WNV forecasting, allowing for more robust early-warning systems.

Among the models tested, Random Forest consistently outperformed Support Vector Machines (SVM) and Histogram-Based Gradient Boosting Regression (HGBR), achieving a Q² close to 0.5 and a Mean Squared Error (MSE) below 13. This suggests that ensemble-based approaches, which capture complex interactions between features, are particularly well-suited for modeling WNV dynamics. The strong predictive influence of temperature, average daylight length, and northward wind aligns with prior ecological findings on *Culex tarsalis* adaptation, reinforcing the role of these environmental drivers in WNV transmission.

This study highlights several key advancements in West Nile Virus (WNV) prediction, demonstrating the advantages of integrating autoregressive modeling with machine learning techniques.

* Enhanced Predictive Accuracy with Autoregression

Incorporating time-lagged case data significantly improved model performance, underscoring the importance of historical trends in forecasting WNV outbreaks. The inclusion of past case trends enabled the models to capture temporal dependencies more effectively, leading to notable improvements in predictive accuracy. Compared to models relying solely on environmental variables, the autoregression-enhanced models demonstrated superior performance, as reflected in increased Q² values and reduced Mean Squared Error (MSE).

* Feature Importance and Ecological Relevance

Analysis of model feature importance revealed that temperature, average daylight length, and northward wind were among the most influential predictors of WNV incidence. These findings align with the known environmental adaptations of *Culex tarsalis*, further reinforcing the biological plausibility of the predictive models. The SHAP analysis (Figure 5-10) provided additional interpretability, offering insights into how both environmental and autoregressive variables contribute to model predictions. By capturing these critical ecological factors, the models offer a biologically meaningful framework for anticipating WNV transmission dynamics.

* Uncertainty Estimation and Model Robustness

Given the inherent variability in disease forecasting, quantifying uncertainty is essential for model reliability. The implementation of bootstrapping techniques (Figure 5-12) enabled the estimation of 95% confidence intervals, providing a measure of uncertainty around predictions. This approach reduces the risk of overconfident forecasts and ensures that the models remain statistically robust, even in cases of limited or noisy data.

Despite these promising advancements, several challenges remain that warrant further investigation:

* Temporal and Concept Drift

The models exhibited lower predictive performance in certain years (e.g., 2018, 2019, 2021), suggesting that unmodeled factors such as changes in vector control efforts, shifts in surveillance practices, or climate anomalies may have influenced disease transmission patterns. These findings highlight the need for adaptive learning frameworks that can dynamically update as new data becomes available, ensuring that the models remain responsive to evolving epidemiological trends.

* Data Quality and Missing Values

Discrepancies between the official CDPH dataset and the independently scraped dataset indicate potential inconsistencies in WNV surveillance data. While imputation techniques were used to address missing values, biases may persist, particularly in under reported cases. Improving data harmonization efforts across surveillance systems will be crucial for enhancing model reliability.

* Computational Complexity and Scalability

While ensemble-based models outperformed other approaches in our autoregression approach, their computational demands were substantially higher compared to simpler linear models. As the volume of surveillance data increases, optimizing model efficiency becomes a priority. Future efforts should explore techniques such as model distillation, feature selection, and distributed computing frameworks to balance predictive power with computational feasibility.

# Conclusion and Future Directions

West Nile Virus (WNV) continues to pose a significant public health challenge, necessitating a comprehensive understanding of its transmission dynamics and predictive modeling approaches. This dissertation integrates ecological, genetic, and epidemiological data to enhance WNV prediction efforts. Through the application of machine learning models and population genetics analyses, this work provides novel insights into the factors influencing WNV outbreaks and the role of *Culex tarsalis* in its transmission.

## Summary of key findings

**Population Genetics of *Culex tarsalis* (Chapter 3)**

* A detailed analysis of the genetic differentiation and adaptation of *Cx. tarsalis* revealed distinct population structures driven by environmental selection pressures.
* Key genetic markers associated with climate variables were identified, highlighting the influence of temperature, precipitation, and wind patterns on mosquito distribution and vector competence.

**National-Level WNV Prediction Model (Chapter 4)**

* Using historical surveillance data and environmental factors, a national machine learning model was developed to predict WNV outbreaks.
* However, results showed that national-scale models struggled to capture the variability in WNV transmission across different regions, leading to inconsistent predictive accuracy.
* The inconsistent data collection and lack of localized environmental and ecological details likely contributed to the model’s limited effectiveness, highlighting the challenges of applying a broad-scale predictive approach to a complex and spatially heterogeneous disease.

**Regional WNV Prediction in California (Chapter 5)**

* A refined regional model for California was developed to assess WNV risk at a more localized scale, incorporating high-resolution environmental and disease surveillance data.
* Findings indicated that regional models provided more reliable predictions compared to the national model, likely due to their ability to capture localized environmental conditions and vector-host interactions more effectively, as well as the availability of more consistent surveillance data.
* The results underscore the importance of tailoring predictive models to specific geographic regions rather than relying on a one-size-fits-all national approach.

## Future directions

**Improving Surveillance Data for More Reliable Predictions**

A more standardized and consistent surveillance methodology is needed to reduce reporting inconsistencies and biases across different regions. Enhanced data collection strategies, including improved case reporting, expanded vector monitoring, and integration of real-time environmental and genomic data, could significantly improve model training and predictive accuracy.

**Integrating Genomic Data into Prediction Models**

Future studies could incorporate west nile virus, mosquito and avian host genomic data to refine vector competence estimates and improve outbreak risk assessments.

**Developing Real-Time Adaptive Models**

Implementing models that continuously update predictions based on incoming surveillance data could improve responsiveness to emerging outbreaks.

**Expanding Regional Models**

Given the success of the California-specific model, similar approaches could be applied to other high-risk regions to refine localized WNV forecasting and intervention strategies if more consistent surveillance data existed.

## Final remarks

This dissertation underscores the importance of integrating multidisciplinary approaches in vector-borne disease modeling within a One Health framework. By bridging insights from genetics, epidemiology, computational modeling, and environmental science, this work contributes to a more comprehensive and holistic framework for WNV prediction and control. The research incorporates diverse and complex data sources—including genomic, climatic, ecological, and epidemiological datasets—to better capture the multifaceted nature of WNV transmission dynamics.

The findings demonstrate that while national-scale models face challenges in accurately predicting WNV outbreaks due to spatial heterogeneity, regional models tailored to local environmental conditions show greater promise. The integration of high-resolution, region-specific data allows for improved forecasting accuracy and more effective public health interventions.

As climate and ecological conditions continue to evolve, adaptive and data-driven strategies rooted in the One Health approach will be essential in mitigating the future impact of WNV and similar arboviral diseases. By leveraging diverse data streams and improving surveillance methodologies, future research can further enhance predictive capabilities and inform targeted control strategies to reduce the burden of vector-borne diseases.

# Reference

1. [Mackenzie JS, Jeggo M. The one health approach-why is it so important? Trop Med Infect Dis. 2019;4: 88. doi:](http://paperpile.com/b/wo71Nx/gyD4)[10.3390/tropicalmed4020088](http://dx.doi.org/10.3390/tropicalmed4020088)

2. [Prata JC, Ribeiro AI, Rocha-Santos T. An introduction to the concept of One Health. One Health. Elsevier; 2022. pp. 1–31. doi:](http://paperpile.com/b/wo71Nx/7ggK)[10.1016/b978-0-12-822793-7.00003-6](http://dx.doi.org/10.1016/b978-0-12-822794-7.00004-6)

3. [Ghai RR, Wallace RM, Kile JC, Shoemaker TR, Vieira AR, Negron ME, et al. A generalizable one health framework for the control of zoonotic diseases. Sci Rep. 2022;12: 8588. doi:](http://paperpile.com/b/wo71Nx/VjF3)[10.1038/s41598-022-12619-1](http://dx.doi.org/10.1038/s41598-022-12619-1)

4. [Rai BD, Tessema GA, Fritschi L, Pereira G. The application of the One Health approach in the management of five major zoonotic diseases using the World Bank domains: A scoping review. One Health. 2024;18: 100695. doi:](http://paperpile.com/b/wo71Nx/fQ2V)[10.1016/j.onehlt.2024.100695](http://dx.doi.org/10.1016/j.onehlt.2024.100695)

5. [Rabinowitz PM, Kock R, Kachani M, Kunkel R, Thomas J, Gilbert J, et al. Toward proof of concept of a one health approach to disease prediction and control. Emerg Infect Dis. 2013;19. doi:](http://paperpile.com/b/wo71Nx/jIgw)[10.3201/eid1912.130265](http://dx.doi.org/10.3201/eid1912.130265)

6. [Leonelli S, Tempini N. Where health and environment meet: the use of invariant parameters in big data analysis. Synthese. 2021;198: 2485–2504. doi:](http://paperpile.com/b/wo71Nx/eUaY)[10.1007/s11229-018-1843-2](http://dx.doi.org/10.1007/s11229-018-1844-2)

7. [Cremin CJ, Dash S, Huang X. Big data: Historic advances and emerging trends in biomedical research. Curr Res Biotechnol. 2022;4: 138–151. doi:](http://paperpile.com/b/wo71Nx/vzEH)[10.1016/j.crbiot.2022.02.004](http://dx.doi.org/10.1016/j.crbiot.2022.02.004)

8. [Elrashedy A, Mousa W, Nayel M, Salama A, Zaghawa A, Elsify A, et al. Advances in bioinformatics and multi-omics integration: transforming viral infectious disease research in veterinary medicine. Virol J. 2025;22: 22. doi:](http://paperpile.com/b/wo71Nx/Jja4)[10.1186/s12984-024-02640-x](http://dx.doi.org/10.1186/s12985-025-02640-x)

9. [Schadt C, Martin S, Carrell A, Fortner A, Hopp D, Jacobson D, et al. An integrated metagenomic, metabolomic and transcriptomic survey of Populus across genotypes and environments. Sci Data. 2024;11: 339. doi:](http://paperpile.com/b/wo71Nx/voMO)[10.1038/s41597-023-03069-7](http://dx.doi.org/10.1038/s41597-024-03069-7)

10. [Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23: 689. doi:](http://paperpile.com/b/wo71Nx/nH7o)[10.1186/s12909-023-04698-z](http://dx.doi.org/10.1186/s12909-023-04698-z)

11. [Zhao AP, Li S, Cao Z, Hu PJ-H, Wang J, Xiang Y, et al. AI for science: Predicting infectious diseases. Journal of Safety Science and Resilience. 2024;5: 130–146. doi:](http://paperpile.com/b/wo71Nx/z6fr)[10.1016/j.jnlssr.2024.02.002](http://dx.doi.org/10.1016/j.jnlssr.2024.02.002)

12. [Javaid M, Haleem A, Pratap Singh R, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks. 2022;3: 58–73. doi:](http://paperpile.com/b/wo71Nx/zWoV)[10.1016/j.ijin.2022.05.002](http://dx.doi.org/10.1016/j.ijin.2022.05.002)

13. [Smirnov A, Liao Y, Fahy E, Subramaniam S, Du X. ADAP-KDB: A spectral knowledgebase for tracking and prioritizing unknown GC-MS spectra in the NIH’s Metabolomics Data Repository. Anal Chem. 2021;93: 12213–12220. doi:](http://paperpile.com/b/wo71Nx/RtKn)[10.1021/acs.analchem.1c00355](http://dx.doi.org/10.1021/acs.analchem.1c00355)

14. [Smirnov A, Liao Y, Du X. Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreening. Metabolites. 2022;12. doi:](http://paperpile.com/b/wo71Nx/4BlE)[10.3390/metabo12060491](http://dx.doi.org/10.3390/metabo12060491)

15. [adap-big.github.io. Github; Available:](http://paperpile.com/b/wo71Nx/sZxJG) <https://github.com/ADAP-BIG/adap-big.github.io>

16. [Borrego SL, Fahrmann J, Datta R, Stringari C, Grapov D, Zeller M, et al. Metabolic changes associated with methionine stress sensitivity in MDA-MB-468 breast cancer cells. Cancer Metab. 2016;4: 9. doi:](http://paperpile.com/b/wo71Nx/TcLeQ)[10.1186/s40170-015-0148-6](http://dx.doi.org/10.1186/s40170-016-0148-6)

17. [Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44: D463–70. doi:](http://paperpile.com/b/wo71Nx/ghC05)[10.1093/nar/gkv1042](http://dx.doi.org/10.1093/nar/gkv1042)

18. [Sinka ME, Bangs MJ, Manguin S, Rubio-Palis Y, Chareonviriyaphap T, Coetzee M, et al. A global map of dominant malaria vectors. Parasit Vectors. 2012;5: 69. doi:](http://paperpile.com/b/wo71Nx/xaAv)[10.1186/1755-3304-4-69](http://dx.doi.org/10.1186/1756-3305-5-69)

19. [Venkatesh A, Jain A, Davies H, Periera L, Maki JN, Gomes E, et al. Hospital-derived antibody profiles of malaria patients in Southwest India. Malar J. 2019;18. doi:](http://paperpile.com/b/wo71Nx/flLJ)[10.1186/s12935-019-2771-5](http://dx.doi.org/10.1186/s12936-019-2771-5)

20. [Awasthi G, Das A. Genetics of chloroquine-resistant malaria: a haplotypic view. Mem Inst Oswaldo Cruz. 2013;108: 947–961. doi:](http://paperpile.com/b/wo71Nx/q6xT)[10.1590/0073-0276130274](http://dx.doi.org/10.1590/0074-0276130274)

21. [Memvanga PB, Nkanga CI. Liposomes for malaria management: the evolution from 1980 to 2020. Malar J. 2021;20: 327. doi:](http://paperpile.com/b/wo71Nx/tDvQ)[10.1186/s12935-021-03858-0](http://dx.doi.org/10.1186/s12936-021-03858-0)

22. [Onen H, Luzala MM, Kigozi S, Sikumbili RM, Muanga C-JK, Zola EN, et al. Mosquito-borne diseases and their control strategies: An overview focused on green synthesized plant-based metallic nanoparticles. Insects. 2023;14: 221. doi:](http://paperpile.com/b/wo71Nx/sYZ8)[10.3390/insects14030221](http://dx.doi.org/10.3390/insects14030221)

23. [Mboussou F, Ndumbi P, Ngom R, Kassamali Z, Ogundiran O, Van Beek J, et al. Infectious disease outbreaks in the African region: overview of events reported to the World Health Organization in 2018 - ERRATUM. Epidemiol Infect. 2019;147: e307. doi:](http://paperpile.com/b/wo71Nx/rrvh)[10.1017/S0950268819002061](http://dx.doi.org/10.1017/S0950268819002061)

24. [Kamgang B, Vazeille M, Yougang AP, Tedjou AN, Wilson-Bahun TA, Mousson L, et al. Potential of Aedes albopictus and Aedes aegypti (Diptera: Culicidae) to transmit yellow fever virus in urban areas in Central Africa. Emerg Microbes Infect. 2019;8: 1636–1641. doi:](http://paperpile.com/b/wo71Nx/HqFC)[10.1080/22221751.2019.1688097](http://dx.doi.org/10.1080/22221751.2019.1688097)

25. [Valentine MJ, Murdock CC, Kelly PJ. Sylvatic cycles of arboviruses in non-human primates. Parasit Vectors. 2019;12. doi:](http://paperpile.com/b/wo71Nx/PZwc)[10.1186/s13071-019-3732-0](http://dx.doi.org/10.1186/s13071-019-3732-0)

26. [Cavalcante KRLJ, Tauil PL. Epidemiol Serv Saude. 2017;26: 617–620. doi:](http://paperpile.com/b/wo71Nx/cuuW)[10.5123/S1679-49742017000300018](http://dx.doi.org/10.5123/S1679-49742017000300018)

27. [Couto-Lima D, Madec Y, Bersot MI, Campos SS, Motta M de A, Santos FBD, et al. Potential risk of re-emergence of urban transmission of Yellow Fever virus in Brazil facilitated by competent Aedes populations. Sci Rep. 2017;7: 4848. doi:](http://paperpile.com/b/wo71Nx/roTc)[10.1038/s41598-017-05185-3](http://dx.doi.org/10.1038/s41598-017-05186-3)

28. [Câmara FP, Gomes ALBB, Carvalho LMF de, Castello LGV. Dynamic behavior of sylvatic yellow fever in Brazil (1953-2008). Rev Soc Bras Med Trop. 2011;44: 297–299. doi:](http://paperpile.com/b/wo71Nx/nk5a)[10.1590/s0037-86822011005000024](http://dx.doi.org/10.1590/s0037-86822011005000024)

29. [Ross RW. The Newala epidemic: III. The virus: isolation, pathogenic properties and relationship to the epidemic. J Hyg (Lond). 1956;54: 177–191. doi:](http://paperpile.com/b/wo71Nx/wbg6)[10.1017/s0022172400044442](http://dx.doi.org/10.1017/s0022172400044442)

30. [Silva LA, Dermody TS. Chikungunya virus: epidemiology, replication, disease mechanisms, and prospective intervention strategies. J Clin Invest. 2017;127: 737–749. doi:](http://paperpile.com/b/wo71Nx/k4ed)[10.1172/jci84417](http://dx.doi.org/10.1172/jci84417)

31. [Benelli G, Mehlhorn H. Declining malaria, rising of dengue and Zika virus: insights for mosquito vector control. Parasitol Res. 2016;115: 1747–1754. doi:](http://paperpile.com/b/wo71Nx/gc0H)[10.1007/s00435-015-4971-z](http://dx.doi.org/10.1007/s00436-016-4971-z)

32. [Saxena SK, Elahi A, Gadugu S, Prasad AK. Zika virus outbreak: an overview of the experimental therapeutics and treatment. Virusdisease. 2016;27: 111–115. doi:](http://paperpile.com/b/wo71Nx/34a1)[10.1007/s13337-015-0307-y](http://dx.doi.org/10.1007/s13337-016-0307-y)

33. [Mencattelli G, Ndione MHD, Rosà R, Marini G, Diagne CT, Diagne MM, et al. Epidemiology of West Nile virus in Africa: An underestimated threat. PLoS Negl Trop Dis. 2022;16: e0010075. doi:](http://paperpile.com/b/wo71Nx/YabW)[10.1371/journal.pntd.0010075](http://dx.doi.org/10.1371/journal.pntd.0010075)

34. [Debiasi RL. West nile virus neuroinvasive disease. Curr Infect Dis Rep. 2011;13: 350–359. doi:](http://paperpile.com/b/wo71Nx/ZCoV)[10.1007/s11908-011-0193-9](http://dx.doi.org/10.1007/s11908-011-0193-9)

35. [Davis LE, DeBiasi R, Goade DE, Haaland KY, Harrington JA, Harnar JB, et al. West Nile virus neuroinvasive disease. Ann Neurol. 2006;60: 286–300. doi:](http://paperpile.com/b/wo71Nx/rJqE)[10.1002/ana.20959](http://dx.doi.org/10.1002/ana.20959)

36. [Hart J Jr, Tillman G, Kraut MA, Chiang H-S, Strain JF, Li Y, et al. West Nile virus neuroinvasive disease: neurological manifestations and prospective longitudinal outcomes. BMC Infect Dis. 2014;14: 248. doi:](http://paperpile.com/b/wo71Nx/GWnH)[10.1186/1471-2333-13-248](http://dx.doi.org/10.1186/1471-2334-14-248)

37. [Santini M, Haberle S, Židovec-Lepej S, Savić V, Kusulja M, Papić N, et al. Severe west Nile Virus Neuroinvasive Disease: Clinical characteristics, short- and long-term outcomes. Pathogens. 2022;11: 52. doi:](http://paperpile.com/b/wo71Nx/b2EX)[10.3390/pathogens11010052](http://dx.doi.org/10.3390/pathogens11010052)

38. [Pavesi A, Tiecco G, Rossi L, Sforza A, Ciccarone A, Compostella F, et al. Inflammatory response associated with West Nile neuroinvasive disease: A systematic review. Viruses. 2024;16: 383. doi:](http://paperpile.com/b/wo71Nx/g9VB)[10.3390/v16030383](http://dx.doi.org/10.3390/v16030383)

39. [Wilson AL, Courtenay O, Kelly-Hope LA, Scott TW, Takken W, Torr SJ, et al. The importance of vector control for the control and elimination of vector-borne diseases. PLoS Negl Trop Dis. 2020;14: e0007831. doi:](http://paperpile.com/b/wo71Nx/f6XB)[10.1371/journal.pntd.0007831](http://dx.doi.org/10.1371/journal.pntd.0007831)

40. [Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control: New Edition. Geneva;](http://paperpile.com/b/wo71Nx/DWxa)

41. [Manikandan S, Mathivanan A, Bora B, Hemaladkshmi P, Abhisubesh V, Poopathi S. A review on vector borne disease transmission: Current strategies of mosquito vector control. Indian J Entomol. 2022; 1–11. doi:](http://paperpile.com/b/wo71Nx/KuO1)[10.55446/ije.2022.593](http://dx.doi.org/10.55446/ije.2022.593)

42. [Shaw WR, Catteruccia F. Vector biology meets disease control: using basic research to fight vector-borne diseases. Nat Microbiol. 2019;4: 20–34. doi:](http://paperpile.com/b/wo71Nx/AWcw)[10.1038/s41563-018-0213-7](http://dx.doi.org/10.1038/s41564-018-0214-7)

43. [Anoopkumar AN, Aneesh EM. A critical assessment of mosquito control and the influence of climate change on mosquito-borne disease epidemics. Environ Dev Sustain. 2022;24: 8900–8929. doi:](http://paperpile.com/b/wo71Nx/Ekct)[10.1007/s10668-021-01792-4](http://dx.doi.org/10.1007/s10668-021-01792-4)

44. [Ozdenerol E, Taff GN, Akkus C. Exploring the spatio-temporal dynamics of reservoir hosts, vectors, and human hosts of West Nile virus: a review of the recent literature. Int J Environ Res Public Health. 2013;10: 5399–5432. doi:](http://paperpile.com/b/wo71Nx/6t59)[10.3390/ijerph10115399](http://dx.doi.org/10.3390/ijerph10115399)

45. [Holcomb KM, Staples JE, Nett RJ, Beard CB, Petersen LR, Benjamin SG, et al. Multi-Model Prediction of West Nile Virus Neuroinvasive Disease With Machine Learning for Identification of Important Regional Climatic Drivers. Geohealth. 2023;7: e2023GH000906. doi:](http://paperpile.com/b/wo71Nx/2cTN)[10.1029/2023GH000906](http://dx.doi.org/10.1029/2023GH000906)

46. [Bauer AM, Guralnick RP, Whitehead SA, Barve N, Allen JM, Campbell LP. Land use predicts proportion of West Nile virus vector‐competent mosquitoes. Ecosphere. 2024;15. doi:](http://paperpile.com/b/wo71Nx/scTxk)[10.1002/ecs2.4771](http://dx.doi.org/10.1002/ecs2.4771)

47. [Wilke ABB, Chase C, Vasquez C, Carvajal A, Medina J, Petrie WD, et al. Urbanization creates diverse aquatic habitats for immature mosquitoes in urban areas. Sci Rep. 2019;9: 15335. doi:](http://paperpile.com/b/wo71Nx/5Nwt8)[10.1038/s41598-019-51787-5](http://dx.doi.org/10.1038/s41598-019-51787-5)

48. [Kolimenakis A, Heinz S, Wilson ML, Winkler V, Yakob L, Michaelakis A, et al. The role of urbanisation in the spread of Aedes mosquitoes and the diseases they transmit—A systematic review. PLoS Negl Trop Dis. 2021;15: e0009631. doi:](http://paperpile.com/b/wo71Nx/QaxYv)[10.1371/journal.pntd.0009631](http://dx.doi.org/10.1371/journal.pntd.0009631)

49. [Kamiya T, Greischar MA, Wadhawan K, Gilbert B, Paaijmans K, Mideo N. Temperature-dependent variation in the extrinsic incubation period elevates the risk of vector-borne disease emergence. Epidemics. 2019;30: 100382. doi:](http://paperpile.com/b/wo71Nx/QhMV)[10.1016/j.epidem.2019.100382](http://dx.doi.org/10.1016/j.epidem.2019.100382)

50. [Ahmed T, Hyder MZ, Liaqat I, Scholz M. Climatic conditions: Conventional and nanotechnology-based methods for the control of mosquito vectors causing human health issues. Int J Environ Res Public Health. 2019;16: 3165. doi:](http://paperpile.com/b/wo71Nx/Ezhw)[10.3390/ijerph16173165](http://dx.doi.org/10.3390/ijerph16173165)

51. [Giesen C, Herrador Z, Fernandez-Martinez B, Figuerola J, Gangoso L, Vazquez A, et al. A systematic review of environmental factors related to WNV circulation in European and Mediterranean countries. One Health. 2023;16: 100478. doi:](http://paperpile.com/b/wo71Nx/xKSt)[10.1016/j.onehlt.2022.100478](http://dx.doi.org/10.1016/j.onehlt.2022.100478)

52. [Vargas Campos CA, García-Pérez S, Figuerola J, Martínez-de la Puente J, Polo I, Rodríguez-de-Fonseca B, et al. Comprehensive analysis of West Nile Virus transmission: Environmental, ecological, and individual factors. An umbrella review. One Health. 2025;20: 100984. doi:](http://paperpile.com/b/wo71Nx/JBjE)[10.1016/j.onehlt.2025.100984](http://dx.doi.org/10.1016/j.onehlt.2025.100984)

53. [Paz S, Semenza JC. Environmental drivers of West Nile fever epidemiology in Europe and Western Asia--a review. Int J Environ Res Public Health. 2013;10: 3543–3562. doi:](http://paperpile.com/b/wo71Nx/rkC0)[10.3390/ijerph10083543](http://dx.doi.org/10.3390/ijerph10083543)

54. [Paz S. Climate change impacts on West Nile virus transmission in a global context. Philos Trans R Soc Lond B Biol Sci. 2015;370: 20130561. doi:](http://paperpile.com/b/wo71Nx/plCQ)[10.1098/rstb.2013.0561](http://dx.doi.org/10.1098/rstb.2013.0561)

55. [Filho WL, Scheday S, Boenecke J, Gogoi A, Maharaj A, Korovou S. Climate Change, health and mosquito-borne diseases: Trends and implications to the Pacific region. Int J Environ Res Public Health. 2019;16: 5114. doi:](http://paperpile.com/b/wo71Nx/ZTcg)[10.3390/ijerph16245114](http://dx.doi.org/10.3390/ijerph16245114)

56. [Humphreys JM, Pelzel-McCluskey AM, Cohnstaedt LW, McGregor BL, Hanley KA, Hudson AR, et al. Integrating Spatiotemporal Epidemiology, Eco-Phylogenetics, and Distributional Ecology to Assess West Nile Disease Risk in Horses. Viruses. 2021;13. doi:](http://paperpile.com/b/wo71Nx/rS46)[10.3390/v13091811](http://dx.doi.org/10.3390/v13091811)

57. [Brown HE, Childs JE, Diuk-Wasser MA, Fish D. Ecological factors associated with West Nile virus transmission, northeastern United States. Emerg Infect Dis. 2008;14: 1539–1545. doi:](http://paperpile.com/b/wo71Nx/XjYR)[10.3201/eid1410.071396](http://dx.doi.org/10.3201/eid1410.071396)

58. [Ma J, Xu N, Xu Y, Huang ZYX, Chen C, Wang YXG. Impacts of urbanization and habitat characteristics on the human risk of West Nile Disease in the United States. Biology (Basel). 2025;14: 224. doi:](http://paperpile.com/b/wo71Nx/m77w)[10.3390/biology14030224](http://dx.doi.org/10.3390/biology14030224)

59. [Ruiz MO, Walker ED, Foster ES, Haramis LD, Kitron UD. Association of West Nile virus illness and urban landscapes in Chicago and Detroit. Int J Health Geogr. 2007;6: 10. doi:](http://paperpile.com/b/wo71Nx/e0tz)[10.1186/1475-072X-5-10](http://dx.doi.org/10.1186/1476-072X-6-10)

60. [Wang X, Zhou G, Zhong D, Li Y, Octaviani S, Shin AT, et al. Impact of underground storm drain systems on larval ecology of Culex and Aedes species in urban environments of Southern California. Sci Rep. 2021;11: 12667. doi:](http://paperpile.com/b/wo71Nx/bsf1)[10.1038/s41598-021-92190-3](http://dx.doi.org/10.1038/s41598-021-92190-3)

61. [The Dark Side of Stormwater Runoff Management: Disease Vectors Associated With Structural BMPs.](http://paperpile.com/b/wo71Nx/NSTO)

62. [Wang H-R, Liu T, Gao X, Wang H-B, Xiao J-H. Impact of climate change on the global circulation of West Nile virus and adaptation responses: a scoping review. Infect Dis Poverty. 2024;13: 38. doi:](http://paperpile.com/b/wo71Nx/rBN9)[10.1186/s40249-023-01207-2](http://dx.doi.org/10.1186/s40249-024-01207-2)

63. [Magallanes S, Llorente F, Ruiz-López MJ, Puente JM la, Ferraguti M, Gutiérrez-López R, et al. Warm winters are associated to more intense West Nile virus circulation in southern Spain. Emerg Microbes Infect. 2024;13: 2348510. doi:](http://paperpile.com/b/wo71Nx/xZ28)[10.1080/22221751.2024.2348510](http://dx.doi.org/10.1080/22221751.2024.2348510)

64. [Bakonyi T, Haussig JM. West Nile virus keeps on moving up in Europe. Euro Surveill. 2020;25. doi:](http://paperpile.com/b/wo71Nx/Cjfm)[10.2807/1560-7917.ES.2020.25.46.2001938](http://dx.doi.org/10.2807/1560-7917.ES.2020.25.46.2001938)

65. [Turell MJ, Dohm DJ, Sardelis MR, Oguinn ML, Andreadis TG, Blow JA. An update on the potential of north American mosquitoes (Diptera: Culicidae) to transmit West Nile Virus. J Med Entomol. 2005;42: 57–62. doi:](http://paperpile.com/b/wo71Nx/VU63r)[10.1093/jmedent/42.1.57](http://dx.doi.org/10.1093/jmedent/42.1.57)

66. [Goddard LB, Roth AE, Reisen WK, Scott TW. Vector competence of California mosquitoes for West Nile virus. Emerg Infect Dis. 2002;8: 1385–1391. doi:](http://paperpile.com/b/wo71Nx/J4uy3)[10.3201/eid0812.020536](http://dx.doi.org/10.3201/eid0812.020536)

67. [Jayaraj VJ, Avoi R, Gopalakrishnan N, Raja DB, Umasa Y. Developing a dengue prediction model based on climate in Tawau, Malaysia. Acta Trop. 2019;197: 105055. doi:](http://paperpile.com/b/wo71Nx/qVBu)[10.1016/j.actatropica.2019.105055](http://dx.doi.org/10.1016/j.actatropica.2019.105055)

68. [de Lima CL, da Silva ACG, Moreno GMM, Cordeiro da Silva C, Musah A, Aldosery A, et al. Temporal and spatiotemporal arboviruses forecasting by machine learning: A systematic review. Front Public Health. 2022;10: 900077. doi:](http://paperpile.com/b/wo71Nx/ttyw)[10.3389/fpubh.2022.900077](http://dx.doi.org/10.3389/fpubh.2022.900077)

69. [Vector-borne diseases: studies in human West Nile Virus and canine Lyme nephritis.](http://paperpile.com/b/wo71Nx/0Iof)

70. [Uelmen JA, Irwin P, Bartlett D, Brown W, Karki S, Ruiz MO, et al. Effects of scale on modeling west Nile virus disease risk. Am J Trop Med Hyg. 2021;104: 151–165. doi:](http://paperpile.com/b/wo71Nx/VFss)[10.4269/ajtmh.20-0416](http://dx.doi.org/10.4269/ajtmh.20-0416)

71. [Myer MH, Johnston JM. Spatiotemporal Bayesian modeling of West Nile virus: Identifying risk of infection in mosquitoes with local-scale predictors. Sci Total Environ. 2019;650: 2818–2829. doi:](http://paperpile.com/b/wo71Nx/Vt8m)[10.1016/j.scitotenv.2018.09.397](http://dx.doi.org/10.1016/j.scitotenv.2018.09.397)

72. [Talbot B, Ludwig A, O’Brien SF, Drews SJ, Ogden NH, Kulkarni MA. Spatial and temporal analysis of West Nile virus infection in mosquito and human populations based on arboviral detection testing data. Sci Rep. 2024;14: 31343. doi:](http://paperpile.com/b/wo71Nx/4pvo)[10.1038/s41598-023-82739-3](http://dx.doi.org/10.1038/s41598-024-82739-3)

73. [Phanitchat T, Zhao B, Haque U, Pientong C, Ekalaksananan T, Aromseree S, et al. Spatial and temporal patterns of dengue incidence in northeastern Thailand 2005-2016. BMC Infect Dis. 2019;19: 743. doi:](http://paperpile.com/b/wo71Nx/t1p4)[10.1186/s12879-019-4379-3](http://dx.doi.org/10.1186/s12879-019-4379-3)

74. [Morin CW, Comrie AC. Regional and seasonal response of a West Nile virus vector to climate change. Proc Natl Acad Sci U S A. 2013;110: 15620–15625. doi:](http://paperpile.com/b/wo71Nx/0YRW)[10.1073/pnas.1307135110](http://dx.doi.org/10.1073/pnas.1307135110)

75. [Hartley DM, Barker CM, Le Menach A, Niu T, Gaff HD, Reisen WK. Effects of temperature on emergence and seasonality of West Nile virus in California. Am J Trop Med Hyg. 2012;86: 884–894. doi:](http://paperpile.com/b/wo71Nx/mmNs)[10.4269/ajtmh.2012.11-0342](http://dx.doi.org/10.4269/ajtmh.2012.11-0342)

76. [Angelou A, Pappa A, Markov PV, Gewehr S, Stilianakis NI, Kioutsioukis I. Early warning system of the seasonal west nile virus infection risk in humans in northern greece, 2020-2024. Sci Rep. 2025;15: 7129. doi:](http://paperpile.com/b/wo71Nx/oxVV)[10.1038/s41598-024-91995-9](http://dx.doi.org/10.1038/s41598-025-91996-9)

77. [Cissé B, Lapen DR, Chalvet-Monfray K, Ogden NH, Ludwig A. Modeling West Nile Virus transmission in birds and humans: Advantages of using a cellular automata approach. Infect Dis Model. 2024;9: 278–297. doi:](http://paperpile.com/b/wo71Nx/vtem)[10.1016/j.idm.2024.01.002](http://dx.doi.org/10.1016/j.idm.2024.01.002)

78. [Brüssow H, Figuerola J. The spread of the mosquito-transmitted west Nile virus in North America and Europe. Microb Biotechnol. 2025;18: e70120. doi:](http://paperpile.com/b/wo71Nx/Vb7c)[10.1111/1751-7915.70120](http://dx.doi.org/10.1111/1751-7915.70120)

79. [Gubler DJ. Resurgent vector-borne diseases as a global health problem. Emerg Infect Dis. 1998;4: 442–450. doi:](http://paperpile.com/b/wo71Nx/j4uom)[10.3201/eid0403.980326](http://dx.doi.org/10.3201/eid0403.980326)

80. [Sutherst RW. Global change and human vulnerability to vector-borne diseases. Clin Microbiol Rev. 2004;17: 136–173. doi:](http://paperpile.com/b/wo71Nx/hhdU6)[10.1128/cmr.17.1.135-173.2004](http://dx.doi.org/10.1128/cmr.17.1.136-173.2004)

81. [Githeko AK, Lindsay SW, Confalonieri UE, Patz JA. Climate change and vector-borne diseases: a regional analysis. Bull World Health Organ. 2000;78: 1136–1147. Available:](http://paperpile.com/b/wo71Nx/q1K32) <https://www.ncbi.nlm.nih.gov/pubmed/11019462>

82. [Darsie RF, Ward RA. Identification and Geographical Distribution of the Mosquitoes of North America, North of Mexico. 2016. Available:](http://paperpile.com/b/wo71Nx/kqVqG) <https://books.google.com/books/about/Identification_and_Geographical_Distribu.html?hl=&id=kL_ejwEACAAJ>

83. [Reeves WC. Epidemiology and Control of Mosquito-borne Arboviruses in California, 1943-1987. 1990. Available:](http://paperpile.com/b/wo71Nx/H0Rzm) <https://books.google.com/books/about/Epidemiology_and_Control_of_Mosquito_bor.html?hl=&id=LqtFAAAAIAAJ>

84. [Evans MV, Dallas TA, Han BA, Murdock CC, Drake JM. Data-driven identification of potential Zika virus vectors. Elife. 2017;6. doi:](http://paperpile.com/b/wo71Nx/wE7w0)[10.7554/eLife.22053](http://dx.doi.org/10.7554/eLife.22053)

85. [Main BJ, Nicholson J, Winokur OC, Steiner C, Riemersma KK, Stuart J, et al. Vector competence of Aedes aegypti, Culex tarsalis, and Culex quinquefasciatus from California for Zika virus. PLoS Negl Trop Dis. 2018;12: e0006524. doi:](http://paperpile.com/b/wo71Nx/iRZhW)[10.1371/journal.pntd.0006524](http://dx.doi.org/10.1371/journal.pntd.0006524)

86. [Venkatesan M, Rasgon JL. Population genetic data suggest a role for mosquito-mediated dispersal of West Nile virus across the western United States. Mol Ecol. 2010;19: 1573–1584. doi:](http://paperpile.com/b/wo71Nx/Y33Bl)[10.1111/j.1364-294X.2010.04577.x](http://dx.doi.org/10.1111/j.1365-294X.2010.04577.x)

87. [Pfeiler E, Flores-López CA, Mada-Vélez JG, Escalante-Verdugo J, Markow TA. Genetic diversity and population genetics of mosquitoes (Diptera: Culicidae: Culex spp.) from the Sonoran Desert of North America. ScientificWorldJournal. 2013;2013: 724609. doi:](http://paperpile.com/b/wo71Nx/booMa)[10.1155/2013/724609](http://dx.doi.org/10.1155/2013/724609)

88. [Venkatesan M, Westbrook CJ, Hauer MC, Rasgon JL. Evidence for a population expansion in the West Nile virus vector Culex tarsalis. Mol Biol Evol. 2007;24: 1208–1218. doi:](http://paperpile.com/b/wo71Nx/nL0aG)[10.1093/molbev/msm040](http://dx.doi.org/10.1093/molbev/msm040)

89. [Diniz DFA, de Albuquerque CMR, Oliva LO, de Melo-Santos MAV, Ayres CFJ. Diapause and quiescence: dormancy mechanisms that contribute to the geographical expansion of mosquitoes and their evolutionary success. Parasit Vectors. 2017;10: 310. doi:](http://paperpile.com/b/wo71Nx/lWmMT)[10.1186/s13071-017-2234-0](http://dx.doi.org/10.1186/s13071-017-2235-0)

90. [Wilke ABB, Vidal PO, Suesdek L, Marrelli MT. Population genetics of neotropical Culex quinquefasciatus (Diptera: Culicidae). Parasit Vectors. 2014;7: 468. doi:](http://paperpile.com/b/wo71Nx/O4Np5)[10.1186/s13071-013-0468-8](http://dx.doi.org/10.1186/s13071-014-0468-8)

91. [Kotsakiozi P, Richardson JB, Pichler V, Favia G, Martins AJ, Urbanelli S, et al. Population genomics of the Asian tiger mosquito, : insights into the recent worldwide invasion. Ecol Evol. 2017;7: 10143–10157. doi:](http://paperpile.com/b/wo71Nx/NmQJU)[10.1002/ece3.3514](http://dx.doi.org/10.1002/ece3.3514)

92. [Cohnstaedt LW, Snyder D, Maki E, Schafer S. Crowdsourcing methodology: establishing the Cervid Disease Network and the North American Mosquito Project. Vet Ital. 2016;52: 195–200. doi:](http://paperpile.com/b/wo71Nx/cF5oS)[10.12834/VetIt.577.2765.2](http://dx.doi.org/10.12834/VetIt.577.2765.2)

93. [Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods. 2017;14: 1072–1074. doi:](http://paperpile.com/b/wo71Nx/QvI3a)[10.1038/nmeth.4432](http://dx.doi.org/10.1038/nmeth.4432)

94. [Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18: 188–196. doi:](http://paperpile.com/b/wo71Nx/qQ8xv)[10.1101/gr.6743907](http://dx.doi.org/10.1101/gr.6743907)

95. [Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 2017;18: 762–776. doi:](http://paperpile.com/b/wo71Nx/Zxq2m)[10.1016/j.celrep.2016.12.063](http://dx.doi.org/10.1016/j.celrep.2016.12.063)

96. [Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240. doi:](http://paperpile.com/b/wo71Nx/VjHQW)[10.1093/bioinformatics/btu031](http://dx.doi.org/10.1093/bioinformatics/btu031)

97. [Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962: 227–245. doi:](http://paperpile.com/b/wo71Nx/y0T1W)[10.1007/978-1-4939-9173-0\_14](http://dx.doi.org/10.1007/978-1-4939-9173-0_14)

98. [Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:](http://paperpile.com/b/wo71Nx/wPrwE)[10.1093/bioinformatics/btt086](http://dx.doi.org/10.1093/bioinformatics/btt086)

99. [Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol. 2011;772: 157–178. doi:](http://paperpile.com/b/wo71Nx/AFmJ7)[10.1007/978-1-61779-228-1\_9](http://dx.doi.org/10.1007/978-1-61779-228-1_9)

100. [Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22: 3124–3140. doi:](http://paperpile.com/b/wo71Nx/g58kA)[10.1111/mec.12354](http://dx.doi.org/10.1111/mec.12354)

101. [Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:](http://paperpile.com/b/wo71Nx/6vy1P)[10.1093/bioinformatics/btp324](http://dx.doi.org/10.1093/bioinformatics/btp324)

102. [Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:](http://paperpile.com/b/wo71Nx/uHjqv)[10.1093/bioinformatics/btp352](http://dx.doi.org/10.1093/bioinformatics/btp352)

103. [McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:](http://paperpile.com/b/wo71Nx/W1Ji2)[10.1101/gr.107524.110](http://dx.doi.org/10.1101/gr.107524.110)

104. [Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:](http://paperpile.com/b/wo71Nx/zqxTr)[10.1093/bioinformatics/btr330](http://dx.doi.org/10.1093/bioinformatics/btr330)

105. [Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11: 1–9. doi:](http://paperpile.com/b/wo71Nx/8BRG9)[10.1038/nprot.2015.123](http://dx.doi.org/10.1038/nprot.2015.123)

106. [Copernicus Climate Change Service. ERA5-Land monthly averaged data from 2001 to present. ECMWF; 2019. doi:](http://paperpile.com/b/wo71Nx/fbT8b)[10.24381/CDS.68D2BB30](http://dx.doi.org/10.24381/CDS.68D2BB30)

107. [Kamvar ZN, López-Uribe MM, Coughlan S, Grünwald NJ, Lapp H, Manel S. Developing educational resources for population genetics in R: an open and collaborative approach. Mol Ecol Resour. 2017;17: 120–128. doi:](http://paperpile.com/b/wo71Nx/TUI6T)[10.1111/1754-0998.12558](http://dx.doi.org/10.1111/1755-0998.12558)

108. [Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:](http://paperpile.com/b/wo71Nx/lSxQQ)[10.1101/gr.094052.109](http://dx.doi.org/10.1101/gr.094052.109)

109. [Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). Measurement . 2019;17: 160–167. doi:](http://paperpile.com/b/wo71Nx/k1A5A)[10.1080/15366367.2019.1565254](http://dx.doi.org/10.1080/15366367.2019.1565254)

110. [Yu G. scatterpie: Scatter pie plot. R package version 01.](http://paperpile.com/b/wo71Nx/3OwuH)

111. [Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992;131: 479–491. doi:](http://paperpile.com/b/wo71Nx/bC957)[10.1093/genetics/131.2.479](http://dx.doi.org/10.1093/genetics/131.2.479)

112. [Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014;2: e281. doi:](http://paperpile.com/b/wo71Nx/Ich55)[10.7717/peerj.281](http://dx.doi.org/10.7717/peerj.281)

113. [Goudet J. hierfstat, a package for r to compute and test hierarchical  *F*‐statistics. Mol Ecol Notes. 2005;5: 184–186. doi:](http://paperpile.com/b/wo71Nx/WAQ2U)[10.1111/j.1471-8286.2004.00828.x](http://dx.doi.org/10.1111/j.1471-8286.2004.00828.x)

114. [Weir BS, Cockerham CC. ESTIMATING F‐STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution . 1984;38: 1358–1370. doi:](http://paperpile.com/b/wo71Nx/pgvwC)[10.1111/j.1558-5646.1984.tb05657.x](http://dx.doi.org/10.1111/j.1558-5646.1984.tb05657.x)

115. [Thioulouse J, Dray S, Dufour A-B, Siberchicot A, Jombart T, Pavoine S. Multivariate Analysis of Ecological Data with ade4. Springer; 2018. Available:](http://paperpile.com/b/wo71Nx/jHh4r) <https://play.google.com/store/books/details?id=UuN3DwAAQBAJ>

116. [Slatkin M. ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. Evolution. 1993;47: 264–279. doi:](http://paperpile.com/b/wo71Nx/fGiXu)[10.1111/j.1558-5646.1993.tb01215.x](http://dx.doi.org/10.1111/j.1558-5646.1993.tb01215.x)

117. [Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014;23: 5649–5662. doi:](http://paperpile.com/b/wo71Nx/R1JLj)[10.1111/mec.12938](http://dx.doi.org/10.1111/mec.12938)

118. [Jiang S, Luo M-X, Gao R-H, Zhang W, Yang Y-Z, Li Y-J, et al. Isolation-by-environment as a driver of genetic differentiation among populations of the only broad-leaved evergreen shrub Ammopiptanthus mongolicus in Asian temperate deserts. Sci Rep. 2019;9: 12008. doi:](http://paperpile.com/b/wo71Nx/ETj2R)[10.1038/s41598-019-48472-y](http://dx.doi.org/10.1038/s41598-019-48472-y)

119. [Sokal RR, Rohlf FJ. Biometry. 3rd edNew York WH Freeman and Company. 1995.](http://paperpile.com/b/wo71Nx/BLwKk)

120. [Wagner HH, Fortin MJ. Basics of spatial data analysis: linking landscape and genetic data for landscape genetic studies. Landscape genetics. 2015. Available:](http://paperpile.com/b/wo71Nx/IDtbY) <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118525258.ch05>

121. [O’brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007;41: 673–690. doi:](http://paperpile.com/b/wo71Nx/TqaU7)[10.1007/s11134-005-9018-6](http://dx.doi.org/10.1007/s11135-006-9018-6)

122. [Gareth J, Daniela W, Trevor H, Robert T. An introduction to statistical learning: with applications in R. Spinger; 2013. Available:](http://paperpile.com/b/wo71Nx/xTOxg) <https://dspace.agu.edu.vn/handle/agu_library/13322>

123. [Frichot E, Schoville SD, Bouchard G, François O. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Mol Biol Evol. 2013;30: 1687–1699. doi:](http://paperpile.com/b/wo71Nx/x6znI)[10.1093/molbev/mst063](http://dx.doi.org/10.1093/molbev/mst063)

124. [Caye K, Jumentier B, Lepeule J, François O. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Mol Biol Evol. 2019;36: 852–860. doi:](http://paperpile.com/b/wo71Nx/2etUs)[10.1093/molbev/msz008](http://dx.doi.org/10.1093/molbev/msz008)

125. [van den Wollenberg AL. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika. 1977;42: 207–219. doi:](http://paperpile.com/b/wo71Nx/8mLeK)[10.1007/BF02294050](http://dx.doi.org/10.1007/BF02294050)

126. [Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. Package “vegan.” 2019 [cited 14 Nov 2023]. Available:](http://paperpile.com/b/wo71Nx/17wPy) <https://cran.r-hub.io/web/packages/vegan/vegan.pdf>

127. [Borcard D, Gillet F, Legendre P. Numerical Ecology with R. Springer New York; doi:](http://paperpile.com/b/wo71Nx/Io1tp)[10.1007/978-1-4419-7975-6](http://dx.doi.org/10.1007/978-1-4419-7976-6)

128. [Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008;180: 977–993. doi:](http://paperpile.com/b/wo71Nx/HDQbj)[10.1534/genetics.108.092221](http://dx.doi.org/10.1534/genetics.108.092221)

129. [Foll M, Fischer MC, Heckel G, Excoffier L. Estimating population structure from AFLP amplification intensity. Mol Ecol. 2010;19: 4638–4647. doi:](http://paperpile.com/b/wo71Nx/9k7am)[10.1111/j.1364-294X.2010.04820.x](http://dx.doi.org/10.1111/j.1365-294X.2010.04820.x)

130. [Fischer MC, Foll M, Excoffier L, Heckel G. Enhanced AFLP genome scans detect local adaptation in high‐altitude populations of a small rodent (Microtus arvalis). Mol Ecol. 2011;20: 1450–1462. doi:](http://paperpile.com/b/wo71Nx/bZlDe)[10.1111/j.1364-294x.2011.05015.x](http://dx.doi.org/10.1111/j.1365-294x.2011.05015.x)

131. [Privé F, Luu K, Vilhjálmsson BJ, Blum MGB. Performing Highly Efficient Genome Scans for Local Adaptation with R Package pcadapt Version 4. Mol Biol Evol. 2020;37: 2153–2154. doi:](http://paperpile.com/b/wo71Nx/r3Tog)[10.1093/molbev/msaa053](http://dx.doi.org/10.1093/molbev/msaa053)

132. [Boes KE, Ribeiro JMC, Wong A, Harrington LC, Wolfner MF, Sirot LK. Identification and characterization of seminal fluid proteins in the Asian tiger mosquito, Aedes albopictus. PLoS Negl Trop Dis. 2014;8: e2946. doi:](http://paperpile.com/b/wo71Nx/G1pzW)[10.1371/journal.pntd.0002946](http://dx.doi.org/10.1371/journal.pntd.0002946)

133. [Amaro IA, Ahmed-Braimah YH, League GP, Pitcher SA, Avila FW, Cruz PC, et al. Seminal fluid proteins induce transcriptome changes in the Aedes aegypti female lower reproductive tract. BMC Genomics. 2021;22: 896. doi:](http://paperpile.com/b/wo71Nx/lyGvQ)[10.1186/s12863-021-08201-0](http://dx.doi.org/10.1186/s12864-021-08201-0)

134. [Nakamura M, Baldwin D, Hannaford S, Palka J, Montell C. Defective proboscis extension response (DPR), a member of the Ig superfamily required for the gustatory response to salt. J Neurosci. 2002;22: 3463–3472. doi:](http://paperpile.com/b/wo71Nx/lYeUP)[10.1523/JNEUROSCI.22-09-03463.2002](http://dx.doi.org/10.1523/JNEUROSCI.22-09-03463.2002)

135. [Jogl G, Hsiao Y-S, Tong L. Structure and function of carnitine acyltransferases. Ann N Y Acad Sci. 2004;1033: 17–29. doi:](http://paperpile.com/b/wo71Nx/tOlnG)[10.1196/annals.1320.002](http://dx.doi.org/10.1196/annals.1320.002)

136. [Meuti ME, Stone M, Ikeno T, Denlinger DL. Functional circadian clock genes are essential for the overwintering diapause of the Northern house mosquito, Culex pipiens. J Exp Biol. 2015;218: 412–422. doi:](http://paperpile.com/b/wo71Nx/TjXw1)[10.1242/jeb.113233](http://dx.doi.org/10.1242/jeb.113233)

137. [Chang V, Meuti ME. Circadian transcription factors differentially regulate features of the adult overwintering diapause in the Northern house mosquito, Culex pipiens. Insect Biochem Mol Biol. 2020;121: 103365. doi:](http://paperpile.com/b/wo71Nx/E0i1J)[10.1016/j.ibmb.2020.103365](http://dx.doi.org/10.1016/j.ibmb.2020.103365)

138. [Shetty V, Meyers JI, Zhang Y, Merlin C, Slotman MA. Impact of disabled circadian clock on yellow fever mosquito Aedes aegypti fitness and behaviors. Sci Rep. 2022;12: 6899. doi:](http://paperpile.com/b/wo71Nx/Q5Fwo)[10.1038/s41598-022-10824-5](http://dx.doi.org/10.1038/s41598-022-10825-5)

139. [Konopka RJ, Benzer S. Clock mutants of Drosophila melanogaster. Proc Natl Acad Sci U S A. 1971;68: 2112–2116. doi:](http://paperpile.com/b/wo71Nx/xVKTz)[10.1073/pnas.68.9.2112](http://dx.doi.org/10.1073/pnas.68.9.2112)

140. [Denlinger DL. Regulation of diapause. Annu Rev Entomol. 2002;47: 93–122. doi:](http://paperpile.com/b/wo71Nx/8rXSi)[10.1146/annurev.ento.47.091201.145137](http://dx.doi.org/10.1146/annurev.ento.47.091201.145137)

141. [Fagre AC, Lyons S, Staples JE, Lindsey N. West Nile virus and other nationally notifiable arboviral diseases - United States, 2021. MMWR Morb Mortal Wkly Rep. 2023;72: 901–906. doi:](http://paperpile.com/b/wo71Nx/uArf)[10.15585/mmwr.mm7234a1](http://dx.doi.org/10.15585/mmwr.mm7234a1)

142. [Hoyer S, Hamman J. xarray: N-D labeled arrays and datasets in Python. Journal of open research software. 2017;5. doi:](http://paperpile.com/b/wo71Nx/5KmDR)[10.5334/JORS.148](http://dx.doi.org/10.5334/JORS.148)

143. [Tuanmu M-N, Jetz W. A global 1‐km consensus land‐cover product for biodiversity and ecosystem modelling. Glob Ecol Biogeogr. 2014;23: 1031–1045. doi:](http://paperpile.com/b/wo71Nx/4QmRI)[10.1111/geb.12182](http://dx.doi.org/10.1111/geb.12182)

144. [Bradski G, Kaehler A, Others. OpenCV. Dr Dobb’s journal of software tools. 2000;3. Available:](http://paperpile.com/b/wo71Nx/cRPRK) <http://roswiki.autolabor.com.cn/attachments/Events(2f)ICRA2010Tutorial/ICRA_2010_OpenCV_Tutorial.pdf>

145. [Kendall MG. A NEW MEASURE OF RANK CORRELATION. Biometrika. 1938;30: 81–93. doi:](http://paperpile.com/b/wo71Nx/SeXz5)[10.1093/biomet/30.1-2.81](http://dx.doi.org/10.1093/biomet/30.1-2.81)

146. [García-Carrasco J-M, Muñoz A-R, Olivero J, Segura M, Real R. Predicting the spatio-temporal spread of West Nile virus in Europe. PLoS Negl Trop Dis. 2021;15: e0009022. doi:](http://paperpile.com/b/wo71Nx/98UX)[10.1371/journal.pntd.0009022](http://dx.doi.org/10.1371/journal.pntd.0009022)

147. [Semenza JC, Tran A, Espinosa L, Sudre B, Domanovic D, Paz S. Climate change projections of West Nile virus infections in Europe: implications for blood safety practices. Environ Health. 2016;15 Suppl 1: 28. doi:](http://paperpile.com/b/wo71Nx/Cn7O7)[10.1186/s12940-015-0104-4](http://dx.doi.org/10.1186/s12940-016-0105-4)

148. [Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297. doi:](http://paperpile.com/b/wo71Nx/lr1FM)[10.1007/BF00994018](http://dx.doi.org/10.1007/BF00994018)

149. [Ho TK. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press; 2002. doi:](http://paperpile.com/b/wo71Nx/Tzj1I)[10.1109/icdar.1995.598994](http://dx.doi.org/10.1109/icdar.1995.598994)

150. [AIML.com. What are the advantages and disadvantages of Random Forest? AIML.com. 2022 [cited 15 Mar 2025]. Available:](http://paperpile.com/b/wo71Nx/uImM) <https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/>

151. [1.4. Support Vector Machines. In: scikit-learn [Internet]. [cited 15 Mar 2025]. Available:](http://paperpile.com/b/wo71Nx/zGuq) <https://scikit-learn.org/stable/modules/svm.html>

152. [What is a Neural Network? In: GeeksforGeeks [Internet]. 17 Jan 2019 [cited 15 Mar 2025]. Available:](http://paperpile.com/b/wo71Nx/G14k) [https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/#](https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/)

153. [Zhang H. On estimation and prediction for spatial generalized linear mixed models. Biometrics. 2002;58: 129–136. doi:](http://paperpile.com/b/wo71Nx/HmsX)[10.1111/j.0005-341x.2002.00129.x](http://dx.doi.org/10.1111/j.0006-341x.2002.00129.x)

154. [Improve SF. LightGBM Histogram-Based Learning. In: GeeksforGeeks [Internet]. 20 Oct 2023 [cited 15 Mar 2025]. Available:](http://paperpile.com/b/wo71Nx/IuTM) <https://www.geeksforgeeks.org/lightgbm-histogram-based-learning/>

155. [Mwiti D. Gradient Boosted Decision Trees [Guide]: a Conceptual Explanation. In: neptune.ai [Internet]. 21 Jul 2022 [cited 15 Mar 2025]. Available:](http://paperpile.com/b/wo71Nx/wBdk) <https://neptune.ai/blog/gradient-boosted-decision-trees-guide>

156. [Bergstra J, Yamins D, Cox D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. Proceedings of the Python in Science Conference. SciPy; 2013. pp. 13–19. doi:](http://paperpile.com/b/wo71Nx/KSXC)[10.25080/majora-8b375194-003](http://dx.doi.org/10.25080/majora-8b375195-003)

157. [Bergstra JS, Bardenet R, Bengio Y. Kégl: Algorithms for Hyper-Parameter Optimization. Advances in Neural Information Processing Systems. 2011. pp. 2546–2554.](http://paperpile.com/b/wo71Nx/if03)

158. [Komar N, Panella NA, Burkhalter KL. Focal amplification and suppression of West Nile virus transmission associated with communal bird roosts in northern Colorado. J Vector Ecol. 2018;43: 220–234. doi:](http://paperpile.com/b/wo71Nx/09CK)[10.1111/jvec.12306](http://dx.doi.org/10.1111/jvec.12306)

159. [Nemeth NM, Oesterle PT, Bowen RA. Humoral immunity to West Nile virus is long-lasting and protective in the house sparrow (Passer domesticus). Am J Trop Med Hyg. 2009;80: 864–869. doi:](http://paperpile.com/b/wo71Nx/mp9d)[10.4269/ajtmh.2009.80.864](http://dx.doi.org/10.4269/ajtmh.2009.80.864)

160. [Reisen WK. Ecology of West Nile virus in North America. Viruses. 2013;5: 2079–2105. doi:](http://paperpile.com/b/wo71Nx/33UQ)[10.3390/v5092079](http://dx.doi.org/10.3390/v5092079)