PLOS ONE

RESEARCH ARTICLE

Seasonal temperatures and hydrological conditions improve the prediction of West Nile virus infection rates in *Culex* mosquitoes and human case counts in New York and Connecticut

Alexander C. Keyel[1,2]*, Oliver Elison Timm[2], P. Bryon Backenson[3], Catharine Prussing[4], Sarah Quinones[2], Kathleen A. McDonough[1,4], Mathias Vuille[2], Jan E. Conn[1], Philip M. Armstrong[5], Theodore G. Andreadis[5], Laura D. Kramer[1]

1 Division of Infectious Disease, Wadsworth Center, New York State Department of Health, Albany, NY, United States of America, **2** Department of Atmospheric and Environmental Sciences, University at Albany, SUNY, Albany, NY, United States of America, **3** Bureau of Communicable Disease Control, New York State Department of Health, Albany, NY, United States of America, **4** Department of Biomedical Sciences, University at Albany, SUNY, Albany, NY, United States of America, **5** Center for Vector Biology & Zoonotic Diseases, Department of Environmental Sciences, The Connecticut Agricultural Experimental Station, New Haven, CT, United States of America

* akeyel@albany.edu

Citation: Keyel AC, Elison Timm O, Backenson PB, Prussing C, Quinones S, McDonough KA, et al. (2019) Seasonal temperatures and hydrological conditions improve the prediction of West Nile virus infection rates in *Culex* mosquitoes and human case counts in New York and Connecticut. PLoS ONE 14(6): e0217854. https://doi.org/10.1371/journal.pone.0217854

Editor: Jeffrey Shaman, Columbia University, UNITED STATES

Received: January 11, 2019

Accepted: May 19, 2019

Published: June 3, 2019

Data Availability Statement: The relevant data are within the paper and its Supporting Information files. The trap-scale mosquito results cannot be shared publicly due to sensitive information by the New York State Department of Health (NYSDOH) and the Connecticut Agricultural Experimental Station (CAES). The trap-scale mosquito data are available upon request by the owners of the data, NYSDOH and CAES. Data requests may be sent to the NYSDOH Bureau of Communicable Diseases at

Abstract

West Nile virus (WNV; *Flaviviridae*: *Flavivirus*) is a widely distributed arthropod-borne virus that has negatively affected human health and animal populations. WNV infection rates of mosquitoes and human cases have been shown to be correlated with climate. However, previous studies have been conducted at a variety of spatial and temporal scales, and the scale-dependence of these relationships has been understudied. We tested the hypothesis that climate variables are important to understand these relationships at all spatial scales. We analyzed the influence of climate on WNV infection rate of mosquitoes and number of human cases in New York and Connecticut using Random Forests, a machine learning technique. During model development, 66 climate-related variables based on temperature, precipitation and soil moisture were tested for predictive skill. We also included 20–21 non-climatic variables to account for known environmental effects (e.g., land cover and human population), surveillance related information (e.g., relative mosquito abundance), and to assess the potential explanatory power of other relevant factors (e.g., presence of wastewater treatment plants). Random forest models were used to identify the most important climate variables for explaining spatial-temporal variation in mosquito infection rates (abbreviated as *MLE*). The results of the cross-validation support our hypothesis that climate variables improve the predictive skill for *MLE* at county- and trap-scales and for human cases at the county-scale. Of the climate-related variables selected, mean minimum temperature from July–September was selected in all analyses, and soil moisture was selected for the mosquito county-scale analysis. Models demonstrated predictive skill, but still over-

BCDC@health.ny.gov and John.Shepard@ct.gov, 203-974-8517 at CAES. The authors did not have any special access or privileges that other interested, qualified researchers would not have.

Funding: This publication was supported by cooperative agreement 1U01CK000509-01, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services. Initial funding supporting this collaboration was provided by the University at Albany Vice President for Research Office. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and under-estimated WNV *MLE* and numbers of human cases. Models at fine spatial scales had lower absolute errors but had greater errors relative to the mean infection rates.

Introduction

West Nile virus (WNV) has caused 46,086 diagnosed cases in the United States, with over 2000 human deaths (1999–2016) [1]. The ecological impacts of WNV have been even more substantial, as WNV has been found in 332 bird species in the United States [2], caused a 45% decline in American Crows, *Corvus brachyrhynchos*, in the United States [3], killed millions of songbirds (e.g., an estimated 29 million Red-eyed Vireos, *Vireo olivaceus*) [4], and contributed to the listing of the Yellow-billed Magpie, *Pica nuttalli*, as 'Near Threatened' [5]. In addition to avian hosts, WNV has been reported from reptiles [6], mammals [7–9], and amphibians [10]. Non-avian impacts have led to substantial economic losses, notably due to infections of horses and farmed alligators [6,11].

In the Northeast, WNV is primarily found in *Culex* mosquitoes, especially *Cx. pipiens* (e.g., [12]). Avian hosts are thought to be responsible for the majority of WNV amplification [13,14], although species vary widely from non-infectious (e.g., Rock Pigeons, *Columba livia* [14]) to superspreaders (e.g., American Robins, *Turdus migratorius* [15]). Climatic conditions may facilitate WNV through 1) increased mosquito abundances (e.g., [16], 2) increased viral replication rates [17–19], and 3) changing the interactions between mosquitoes and their hosts. Some of these changes could be indirect, such as by affecting timing of breeding or migration for key amplifying species [20,21].

Prior studies have supported this link between WNV and climatic variables [17,22,23]. Prior studies found higher WNV infection rates with increasing temperature [22–27], including winter temperature [23], and higher infection rates with increased growing degree days [28]. Precipitation relationships have been more complex and not consistently detected [27]. Increased precipitation in the preceding year [29] and decreased current year precipitation [24] have been associated with increased WNV transmission. Precipitation may interact with temperature, as drought has been found to be important in WNV dynamics [30,31]. Warm and dry conditions during early spring have been associated with increased WNV activity [22], but not conclusively, as a range of other climate indicators, such as anomalously wet conditions in March, were also identified as potentially important in the same study [22].

Direct comparisons among studies are complicated, as studies have differed in their selection of climate variables, computation of the climate variables, inclusion of important covariates, spatial extent, and spatial resolution (see Table 1). Further, relationships between climate and disease may vary depending on geographic location (e.g., [31]). This has complicated the use of climatic information to produce robust spatial and temporal predictions of WNV prevalence. Differences in analysis extent and resolution (i.e. the scale of the analysis) have been shown to affect analysis results (the modifiable area unit problem, MAUP) [32–34]. One scale-dependent result identified from species distribution models has been that specific relationships with climate at broad spatial scales using aggregated data exist; in contrast, at finer scales other variables may dominate [35]. Indeed, this result has been observed for WNV relative to 30-year climate averages [36]. However, as the life-cycles of mosquitoes and WNV are highly temperature dependent [18,37], strong relationships with variables such as temperature and precipitation may be present even at fine spatial scales.

In this study the main goals were to explore the spatial and temporal relationships between climate and WNV and develop a well-validated statistical model that included climatic as well

Table 1. A summary of literature that includes Connecticut or New York as part of the study area. Studies varied in their choice of dependent variable (De). Independent variables were classified as Surveillance (Su), climate (Cl), land cover (La), Sociological (So), host-related (Ho), or Other (Ot).

Study	Spatial Extent[1]	Spatial Resolution	Temporal Extent	Temporal Resolution	De[3]	Su[4]	Cl[5]	La[7]	So[8]	Ho[9]	Ot[10]
Allan et al. 2009 [38]	USA	County	2002–2004	Annual	H_i	0	0	0	1	2	0
Andreadis et al. 2004 [39]	CT	Point	1999–2003	Annual	H_c	1	0[6]	0	0	0	0
Andreadis et al. 2004 [39]	CT	Point	1999–2003	Annual	M_a	0	0[6]	0	1	0	0
Bowden et al. 2011 [40]	USA	County	2002–2008	7-year period[2]	H_i	0	0	14	0	0	0
Brown et al. 2008a [41]	New Haven, CT	Point	2004	Annual	M_a	0	0	2	0	0	0
Brown et al. 2008b [42]	CT, DE, MA, MD, NJ, NY, PA, RI	County	1999–2006	Annual	H_i	0	0	2	1	0	1
Brownstein et al. 2002 [43]	NY (7 counties)	Point	1999	Annual	H_c	0	0	1	0	0	0
DeFelice et al. 2017 [44]	Suffolk, NY	County	2001–2014	Weekly	H_c	2	0	0	0	0	0
DeFelice et al. 2017 [44]	Suffolk, NY	County	2001–2014	Weekly	M_{MLE}	2	0	0	0	0	0
DeFelice et al. 2018 [45]	USA (12 counties)	County	2001–2016	Weekly	H_c	2	6	0	0	0	0
DeFelice et al. 2018 [45]	USA (12 counties)	County	2001–2016	Weekly	M_{MLE}	2	6	0	0	0	0
Diuk-Wasser et al. 2006 [46]	Fairfield, CT	Point	2001–2003	3-year-period	M_a	0	0	97	1	0	0
Gates and Boston 2009 [47]	USA	County	2004–2006	3-year period	H_c	0	0	1	1	0	0
Gates and Boston 2009 [47]	USA	County	2004–2006	3-year period	E_c	0	0	1	0	0	1
Hahn et al. 2015 [24]	USA	County	2004–2012	Annual	H_{cz}	0	10	0	0	0	0
Keyel et al. (this study)	NY, CT	County, Point	2000–2015	Annual	M_{MLE}	5	66	4	2	7	2
Keyel et al. (this study)	NY, CT	County, Point	2000–2015	Annual	H_c	6	66	4	2	7	2
Landesman et al. 2007 [29]	USA	County	2002–2004	Annual; Monthly	H_c	0	6	0	0	0	0
Little et al. 2016 [22]	Suffolk, NY	13 × 13 km cells	2001–2015	Monthly	M_{MLE}	0	48	0	0	0	0
Liu et al. 2009 [26]	CT	Township	2000–2005	Daily	H_c	5	3	6	1	0	0
Manore et al. 2014 [23]	USA	County	2005–2011	Annual	H_c	2	96	1	6	27	0
Myer et al. 2017 [48]	Suffolk, NY	Point	2008–2014	Weekly	M_{pa}	0	6	37	1	0	0
Myer and Johnston 2019 [49]	Nassau, NY	Point	2001–2015	Weekly	M_{pa}	1	6	16	4	0	2
Paull et al. 2017 [31]	USA	State	1999–2009	Annual	H_{ni}	1	4	0	0	0	0
Rochlin et al. 2008 [50]	Suffolk, NY	Point	2000–2004	Annual	M_a	0	0	10	1	0	1
Rochlin et al. 2008 [50]	Suffolk, NY	Point	2000–2004	Annual	M_p	0	0	10	1	0	1
Rochlin et al. 2009 [51]	Suffolk, NY	Point	1999–2006	Annual	M_a	0	0	3	0	0	2
Rochlin et al. 2011 [52]	Suffolk, NY	Point	2001–2004	4-year period	H_c	8	0	30	13	0	5
Shaman et al. 2011 [25]	Suffolk, NY	13 × 13 km cells	2001–2009	Annual	$M_{\%}$	0	60	0	0	0	0
Tonjes 2008 [53]	Suffolk, NY	Zip Codes	2000–2004	Annual	H_c	2	0	0	1	0	0
Trawinski and MacKay 2008 [28]	Erie, NY	Point	2001–2005	Weekly	M_a	0	33	0	0	0	0
Trawinski and MacKay 2010 [54]	Amherst, Erie, NY	Point	Not reported	2–5 weeks	M_a	0	12	66	27	0	51
Walsh 2012 [27]	NY	County	2000–2010	Annual	H_{pa}	1	2	1	0	0	0
Young et al. 2013 [55]	USA	County	2003–2008	6-year period	H_i	0	30	17	0	0	3

[1] USA: United States of America; CT: Connecticut; DE: Delaware; MA: Massachusetts; MD: Maryland; NJ: New Jersey; NY: New York State; PA: Pennsylvania; RI: Rhode Island.

[2] Assumed. Whether years were pooled or analyzed individually was not clear from the methods section.

[3] *De*: Dependent variables: E_c equine cases; H_c Human cases; H_{cz} z-score deviation from mean number of human cases; H_i Human per-capita incidence; H_{ni} Human West Nile neuroinvasive disease cases only; H_{pa} human cases present or absent; $M_{\%}$ percent of mosquito pools testing positive; M_a Mosquito abundance; M_{MLE} Mosquito infection rate; M_p The proportion of mosquitoes belonging to a particular species, M_{pa} Presence/absence of WNV in mosquito pools.

[4] *Su*: Surveillance variables such as number of dead birds, WNV positive birds, Human WN in previous years, Human infection rate, human immunity (estimated), mosquito infection rates from previous timepoints, mosquito abundance, absence of mosquito surveillance, site classification based on previous WNV infection rates (high, medium, low), number of complaints about mosquitoes, number of known larval sites, WNV positive mosquito pools, distance to nearest complaint, distance to nearest known larval site, distance to nearest WNV positive bird, distance to nearest WNV positive mosquito pool.

[5] *Cl*: Climate and hydrological variables such as temperature, precipitation, growing degree days, and anomalies for each of these variables. Often calculated as minimum, mean, maximum, or cumulative values for different time periods (e.g., month, season, year).

[6] Temperature and rainfall values were discussed, but not statistically related to the WNV results.

[7] *La*: Land cover variables such as percent/proportion land cover for different land cover types, buffer distances, or administrative units or distance to land cover features. Soil drainage characteristics were also included here, as were Normalized Vegetation Difference Index (NDVI), Disease Water Stress Index (DWSI), and Middle Infrared Band.

[8] *So*: Sociological variables such as age (median), education, employment (percent), household income (median), housing age, human population (density), human population (total), race, senior households (count, >65), septic systems (count), vacant housing (percent), urban or rural (categorical).

[9] *Ho*: Host variables such as avian abundance (e.g., by order or species), avian diversity, and community competence.

[10] *Ot*: Other variables, such as aspect, catch basin area, catch basin count, county area, elevation, equine density, flood zone, flood zone (distance to nearest), road length, road polygons (index of fragmentation), slope, wastewater treatment plants (distance from, count per administrative unit), year.

https://doi.org/10.1371/journal.pone.0217854.t001

as non-climatic environmental factors for the northeastern United States. We hypothesized that climatic variables would be important at both coarse (county) and fine (point) spatial scales. Conversely, WNV is widely distributed across many different climatic regions [24,56], and therefore we tested the alternative hypothesis that WNV prevalence and human cases do not depend on climatic variables, and that previous results were due to an omitted, correlated covariate.

Methods & data

Overview

We fit models at two spatial scales, with and without climate variables, and examined the error for a new year of data. We examined statistical relationships between two dependent variables, WNV mosquito infection rates (*MLE)* and human cases of WNV (see *Dependent variables* section), and 86–87 independent variables, grouped into *climate* (66 covariates), *surveillance* (5–6 covariates), host (7 covariates), human population (2 covariates), land cover (4 covariates), and wastewater treatment (2 covariates). Human population, land cover, and wastewater covariates were snapshots in time and were treated as constant across time. Dependent variables were related to independent variables using a random forest analysis [57]. The analysis was conducted with data aggregated over entire years. For *MLE*, relationships were evaluated at two spatial scales (county and trap, see *Scales of analysis*, below), while for human data, due to data restrictions related to privacy concerns, only the county-scale data were used. We used a leave-one-year-out cross validation approach, where the random forest model was fitted using data from all years except one, and the resulting model was used to predict the remaining year. See *Statistical Approach* below for full details. Data processing was performed using Python 2.7 [58], ArcGIS 10.6 (ESRI, Redlands, CA), and R 3.4.3 [59]. All statistical analyses were performed in R [59]. Descriptive information for non-categorical covariates are included in S1 File and a Data Dictionary describing the variables in S2 File. Text files used to run the analyses at the county scales are includes in zipped format as S3 File. For trap-scale data, contact the New York State Department of Health [60] and the Connecticut Agricultral Experimental Station [61] as the trap locations contain sensitive information.

Scales of analysis

Data were analyzed at two scales: aggregated by county (hereafter the county-scale) and from individual mosquito trap locations (hereafter the trap-scale). Due to high uncertainty in some *MLE* for some trap locations (see *Dependent variables* below for *MLE* calculations), trap-scale data were analyzed in two ways, 1) including *MLE* with large confidence intervals, and 2) excluding *MLE* where the estimated 95% confidence interval exceeded 15 infected mosquitoes per 1000 (hereafter the trap-scale subset). Human cases were also analyzed at two different extents: all counties in both New York and Connecticut (hereafter: human all counties) and for just those counties for which mosquito surveillance data were available (hereafter: human subset).

Dependent variables

Human cases. Human case data, aggregated by county and year, were obtained for the entire state of New York (NY) for 2003–2015 [62] and for the state of Connecticut (CT) from 2000–2015 [63]. Data from both states were pooled for the analysis as results were qualitatively similar when each state was analyzed separately (results not shown). Cases of West Nile Fever

and West Nile Neuroinvasive Disease were pooled to increase the sample size, as these two manifestations are highly correlated [40] and this approach has been found to have greater prediction accuracy [55]. West Nile Fever corresponds to clinical cases where the symptoms include fever [64], but the cases were not neuroinvasive. West Nile Neuroinvasive Disease included cases of meningitis, encephalitis, and meningoencephalitis [64]. We note that mild cases of West Nile Fever may go unreported, as the majority of human WNV infections (~80% [65]) do not cause any detectable symptoms, and <1% are neuroinvasive [65,66]. Therefore, the reported cases represent a very small fraction of the human WNV infections. Further, case locations correspond to the patients' county of residence and may not indicate the county where the disease was contracted.

We used total cases as our dependent variable instead of incidence as this approach makes no assumptions about the relationship between West Nile virus cases and total human population. Humans are dead-end hosts for West Nile virus, and therefore do not amplify the virus. As a consequence, the contact rate between mosquitoes and humans is far more important than total human population for determining cases, and this contact rate varies non-linearly with human population. We present a subset of our analyses using incidence in S4 File for comparison purposes.

Mosquito infection rates. *MLE* were calculated based on mosquito trap data. We obtained data from 8 counties in CT [61] and 8 counties in NY State (one western, two central and five southeastern counties) [60]. The data cover the years 2000–2015 (not all data available for all years). These data were pooled from several different mosquito control programs, and each agency employed a slightly different sampling design (see S5 File). For our analysis, any mosquito pool from a CDC Gravid trap [67] that was deployed for less than 24 hours and was tested for WNV was included in the analysis. We restricted the analysis to *Culex pipiens*, *Cx. restuans* and *Cx. salinarius*, as empirical data have demonstrated that *Cx. pipiens* and *Cx. restuans* can amplify WNV [68–71] and *Cx. pipiens* and *Cx. salinarius* can serve as bridge vectors to humans [12,71–73], especially in the northeast [12,39,73,74]. *Cx. pipiens*, *Cx. restuans* and *Cx. salinarius* were pooled in the analysis, because the NY mosquito sampling protocol does not distinguish between *Cx. pipiens* and *Cx. restuans* due to issues with species identification [75]. Mosquito identifications were based on one or more standard references [76–80]. WNV *MLE* were calculated using Maximum Likelihood Estimates [81,82] in R [59,83]. Maximum Likelihood Estimates calculate a mean infection rate and 95% confidence intervals based on the distribution of positive mosquito pools and the number of mosquitoes in each pool and represents a substantial improvement over estimates based on minimum infection rate [81]. We note that the estimates obtained using R for some samples deviated from the estimates obtained with the standard CDC Excel plugin [84], likely due to the omission of a bias correction term in the R version. However, we viewed the magnitude of these inconsistencies as relatively minor compared to the increased convenience of computing the infection rates using R. Samples were pooled for each year based on spatial location. At the trap-scale, 200 m buffers were generated surrounding all trap locations, and any overlapping buffers were treated as a single trap location. This was necessary to correct for minor inconsistencies in the reporting of trap locations across years (e.g., the same trap location may have had a new GPS point collected each year, and in some cases this point may have corresponded to the actual trap location, whereas in other cases this may have corresponded to the center of the area being sampled by the trap). A visual inspection suggested that the use of 200 m buffers (potentially merging traps 400 m apart) adequately addressed these issues, while still maintaining spatial proximity to the original locations.

Covariates

Climate (66 covariates). Climate data were derived from gridded ensemble estimates of daily temperatures and precipitation at 1/8˚ resolution (~12 km × 12 km) for all years included in this study (2000–2015) [85]. The data are available at http://dx.doi.org/10.5065/D6TH8JR2. The gridded data are based on the Global Historical Climatology Network-Daily dataset (GHCN-Daily [86,87]) using daily precipitation and temperature data, with supplemental data from the meteorological observations from the U.S. Natural Resources Conservation Service (NRCS) Snowpack Telemetry (SNOTEL). Daily average temperature and daily temperature range (daily maximum–daily minimum) were interpolated applying a distance-weighted station averaging model. Over the Continental US (CONUS) domain a total of 12,153 (8953) stations provided precipitation (temperature) observations [85]. Precipitation was processed using a similar distance-weighted averaging method. In this particular method the interpolation of precipitation is divided into two components (a) the probability of precipitation (PoP) and (b) the precipitation amount. Furthermore, the method applies an ensemble (therefore probabilistic) interpolation approach, which accounts for the residual variance. This improves the representation of local extremes compared with other gridded daily temperature or precipitation data products. For more details see [85,88].

For the county-scale, climate data were extracted for the county centroids (see S6 File for a justification of the use of the centroid). At the trap-scale, climate data were extracted for the centroid of the merged trap buffers (see *Dependent variables* above, typically the trap location). We aggregated the climate data into four quarters (January–March; April–June; July–September; and October–December). While the majority of WNV cases occur from July to September, we included early season data to account for processes related to the emergence and amplification of WNV, while the late season variables were included to address the end of the WNV season. For each quarter, we calculated the cumulative growing degree days (relative to 10˚C), cumulative precipitation (mm), average precipitation intensity (mm day^{-1}), minimum daily temperature (˚C), maximum daily temperature (˚C), mean minimum temperature (˚C), mean maximum temperature (˚C), and diurnal temperature range (˚C). Minimum and maximum daily temperature correspond to the lowest/highest record observed on a day within the period, while mean minimum and maximum temperature correspond to the average minimum/maximum temperature for the entire period. Growing degree days for quarter 1 were omitted as there were not many days above 10˚C during this three-month period leading to very little variation in this variable. We note that much of the fourth quarter corresponds to the time period after the mosquito season was effectively over, but it was included to capture any effects related to the ending or possible extension of the time when mosquitoes were active.

Growing degree days were calculated as the cumulative number of degree days above 10˚C within the 3-month quarter. Specifically, for each day, 10˚C was subtracted from the mean temperature. If this reduced the value to below zero, zero was used instead, and the sum of all these values for each quarter was computed. We chose 10˚C, as this temperature limit was used in one prior study in the region [26], although a similar rationale could have been used to select 15˚C [28,74]. However, we expected that data from either of these two thresholds would be very highly correlated, so we did not explore the 15˚C threshold. Precipitation intensity was calculated as the total precipitation for each quarter divided by the number of days it rained in that quarter [89]. A minimum of 0.254 mm of rainfall, the threshold of common instrument measurements, was required to count as a rain day.

Daily soil moisture was taken from the NLDAS Soil Moisture 0–200 cm soil depth [90–93], and aggregated to quarterly averages and quarterly means. These were used to calculate

drought anomalies. The quarterly averages were extracted for county centroids and trap locations as above.

Climate anomalies were calculated for each climate variable with respect to a baseline mean for the study period (2000–2015). Although the World Meteorological Organization 30-year baseline is defined as 1981–2010, the use of the study period as a baseline leads to a mean anomaly of zero, which improves the interpretability of the results, and prior studies comparing a study mean to an alternative time period found no meaningful difference [23]. For temperature, the anomaly was calculated as deviation from the mean (Eq 1), whereas for precipitation and drought the anomaly was calculated as percent deviation of the mean (Eqs 2 and 3).

$$T_{anomaly} = T_{quarterly} - \bar{T} \tag{1}$$

$$P_{anomaly} = \frac{(P_{quarterly} - \bar{P})}{\bar{P}} \times 100 \tag{2}$$

$$D_{anomaly} = \frac{(D_{quarterly} - \bar{D})}{\bar{D}} \times 100 \tag{3}$$

T, *P*, and *D* correspond to temperature, precipitation, and drought respectively, and the subscript anomaly refers to the anomaly values for a given year, quarterly refers to the quarterly value for that year, and the bar indicates the quarterly mean across all years.

Surveillance (5–6 covariates). We included trapped mosquito abundance to control for any effects of mosquito population size. Unfortunately, some heterogeneity in this variable exists, i.e., in CT, all mosquitoes captured were tested for WNV, whereas in NY, a maximum of 90 pools of up to 50 mosquitoes were tested (see S5 File for more details). Consequently, the NY data may underestimate the true sampled abundance at the trap. We included abundance divided by the number of pools tested (hereafter called density), to control for unequal sampling efforts across counties and years. The bait used in the CT gravid traps changed from a grass/sod infusion [94] in 2000 and 2001 to a rabbit chow infusion (Purina Mills LLC, St. Louis, MO) for 2003–2005 [39], and was switched to a hay/yeast/lactalbumin infusion [95] starting in 2006. Baits were unspecified for NY counties, with the exception of Suffolk, where a rabbit chow infusion was used [25]. Consequently, a BAIT covariate was added to the analysis as a factor with four levels: unspecified, grass/sod, rabbit chow, and hay/yeast/lactalbumin. The first level, unspecified, was used as a reference level, and was not counted towards the number of covariates. *MLE*, the dependent variable in the mosquito analyses, was included as an independent variable in the human subset analysis.

Host (7 covariates). West Nile virus is an enzootic virus, and is mainly amplified by birds [70]. We obtained avian abundance information from the Breeding Bird Survey (BBS) for each state by year [96]. We included five species for which we had data, that have been identified as especially important in the transmission of WNV: American Robins (*Turdus migratorious*) [15,70], Blue Jays (*Cyanocitta cristata*) [97], Northern Cardinals (*Cardinalis cardinalis*) [70,97], House Sparrows (*Passer domesticus*) [15,70,97] and American Crows (*Corvus brachyrhynchos*) [14]. In addition, we included two aggregated covariates: the total avian abundance and the host reservoir competence index, weighted by abundance. The weighted host competence index (*Eq 4*) was calculated by taking each species' abundance (*a*) multiplied by its host reservoir competence index value. Host reservoir competence index values are the product of susceptibility (*s*, the proportion of birds that become infected as a result of exposure), mean daily infectiousness (*i*, the proportion of exposed vectors that become infectious per day), and

duration of infection (d, the number of days that a bird maintains an infectious viremia) [14,98]. This approach is similar to the approach taken by Kilpatrick et al. [15], but we omitted the feeding preference term due to lack of data.

$$C_w = a * s * i * d \quad (4)$$

The host reservoir competence indices were extracted from the literature [14,97]. Any species without host competence information was excluded. The species in the host competence index included on average 57% percent of the total species abundance. We note that the BBS data are limited, as these data were estimated at the state level (in contrast to Manore et al. [23], who used route-level data to estimate abundances for individual counties). The point count survey method employed by the BBS has also been critiqued on statistical grounds (e.g., [99]).

Human population (2 covariates). Based on previous research, human population is associated with both human cases and *MLE* [39]. Total human population based on the 2010 US Census [100] was obtained for counties (county-scale) and the census tracts containing trap centroids. Human population was converted to population density by dividing by county or tract land area.

Land cover (4 covariates). Based on the results of a prior study [101], we examined the proportion of urban, forest, open, and wetland land cover within each county (county-scale) or within 1000 m of the trap site (trap-scale) using the National Land Cover Data 2011 [102]. A 1000 m buffer was previously identified as being consistently associated with land cover variables [54], but see [46].

Wastewater treatment (2 covariates). Locations of wastewater treatment facilities were obtained from the United States Environmental Protection Agency [103]. These were classified as major or minor. At the county-scale, the number of major facilities and the total number of facilities (highly collinear with the number of minor facilities) were included in the analysis. At the trap-scale, the distance to the nearest major wastewater treatment facility and the distance to any wastewater treatment facility were included in the analysis.

Statistical approach

Correlations were calculated for each of the spatial scales (S7 File). Although correlations among covariates varied substantially (e.g., mean $r_p = 0.08$; 0.00005–0.99 min–max; mosquito infection rates, county scale), no variables were excluded on this basis.

We chose to analyze our data using random forest models [57,104], as preliminary analyses showed random forest approaches had similar or better predictive capability when compared to linear methods (GLMs). Random forest methods take a sample of the data set (with replacement) and construct a cartographic regression tree based on the sample. This approach was then repeated many times (see numbers of trees below) and the final results were obtained by averaging across all trees. Variable importance was assessed using permutation approaches that randomly changed input variables and examined the magnitude of the change in the resulting predictions [57]. For each random forest analysis, the number of variables to try at each split (m) was reached by trying each combination and using the value that corresponded to the best R^2 value. We used 500 trees for screening for m values, and 5000 trees for the final analyses. Each tree had a single terminal node. We evaluated model fit (see *model fit statistics* below for definitions) by examining 1) the root mean squared error (RMSE), 2) RMSE scaled by the mean value, 3) the coefficient of determination, R^2, or the percent of variation explained by the model in the validation set, [105], 4) the Spearman Rank correlation coefficient (r_s) between the predicted and observed values, and 5) the Pearson correlation coefficient (r_p) between the predicted and observed values.

It has been recommended to run random forest models twice: once with all variables of interest, and a second time with a subset of the best variables in order to refine the fit for those variables [57]. Variables with importance scores (see above) greater than the mean importance score were retained in the second pass. Model fit was evaluated using leave-one-year-out cross-validation [106]. The data set was split to omit a single year (and not merely a single observation). A new random forest model was then fitted and the model performance was evaluated for the omitted year. This was repeated for all years. The average skill score was then derived for the leave-one-year-out cross-validation approach. We then assessed the amount of variation uniquely explained by each variable retained in the model based on the cross-validation data set. We then attempted to refine the model further by removing all variables that uniquely contributed less than a threshold value (thresholds tried were 0, 0.001, 0.005, and 0.01) and re-fitting the random forest model. The model corresponding to the threshold that resulted in the highest cross-validated R^2 was then retained and the reported final fit metrics correspond to these final models.

After running the model including all variables of interest, we repeated the analysis approach, but without the climate variables, to identify the degree to which climate variables uniquely contribute to the model fit (variance partitioning, [107,108]). We then repeated this analysis with only the climate variables, to assess the degree to which non-climatic variables influence the choice of climate variables included in the models.

For human case data, we first analyzed the data for the human subset (those counties for which we had both mosquito and human data). We constructed a random forest model to describe the relationships between human cases, the observed *MLE*, the climate covariates, and the other covariates, and partitioned the amount of variance due to each set of covariates [107,108], as was done for mosquitoes above. We then repeated the analysis for the human all counties data set, omitting the *MLE*, again partitioning the amount of variance due to each set of covariates. We note that human cases are discrete occurrences, but the model predictions were on a continuous scale. If this is of concern, we note that the model predictions could be rounded to the nearest integer value.

Model fit statistics

Root Mean Squared Error (RMSE). This was calculated with the RMSE function in the package caret in R [109]. Root mean squared error corresponds to the standard deviation of the residuals [105], and gives an estimate of the magnitude of the errors. It is expressed as number of infected mosquitoes per 1000 mosquitoes or number of human cases.

Median RMSE. RMSE was calculated for each year of the cross-validation data, and median RMSE corresponds to the median RMSE value from this evaluation. Median RMSE is less biased by a single extremely poor or extremely good prediction year.

Scaled RMSE. The Root Mean Squared Error was divided by the mean infection rate, to express the error as a percentage of the mean value. This has the benefit of placing it in the context of the values to be predicted, and may serve as a more intuitive measure of error. This quantity can vary from 0 to ∞, with 0 serving as no error, 1 indicating the model error is equal to the mean, and every value >1 indicating the number of times the error is greater than the mean.

Max error. The maximum error observed for any sample in the validation data set, expressed as number of infected mosquitoes per 1000 mosquitoes, or number of human cases.

Coefficient of determination (R^2). Here, R^2 is defined by *Eq 5*:

$$R^2 = \frac{1 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \quad (5)$$

Where y corresponds to the observed values in the validation set, $\hat{y}$ corresponds to the predicted values, and $\bar{y}$ corresponds to the mean of the validation set [105,110]. In contrast to computing an R^2 for the data used to fit the model, where R^2 is bounded between 0 and 1, by computing R^2 for the validation data set, values can range from $-\infty$ to 1. This occurs because it is possible for the model to have worse predictive power than the mean of the validation data set. A value of 1 would indicate a perfect fit, whereas a value of -1 would correspond to the model having twice the residual squared error of the validation data set's mean value.

Spearman (r_s) and Pearson (r_p) correlation coefficients. These correspond to the Spearman correlation coefficient and the Pearson correlation coefficient, respectively, and were calculated with the cor function in Base R according to the standard formulae [59].

Contour plot methods

We visualized the outputs of the random forest models by creating bivariate contour plots. Model predictions were generated for a regular grid of 100 points covering the parameter space of the two variables. Values for other covariates were fixed at their mean values. Contours were then drawn to indicated lines of equal predicted infection rates. Observed infection rates were overlaid as red circles, with the size proportional to infection rate. However, the observed and predicted infection rates are not strictly comparable, since the observed values are of course affected by state of the other covariates, rather than the mean covariate values. Nonetheless, the contour plots give some insight of the response function that are otherwise hidden in the high-dimensional nonlinear random forest regression model.

Results

We present first a summary of the random forest model fitting results followed by results that address the question of which–if any—climate covariates improve the model skill. In the last part we describe the spatial and temporal variability for selected dependent and independent variables identified by the models as important.

Our best-fitting model using the full data set at the county-scale explained 45% of the variation in *MLE* (Fig 1, Table 2), and 72% of the variation in human case counts (Fig 2, Table 3). When infection rates or case counts were converted to categories, the model both over- and under-predicted WNV risk for some locations and years (Figs 3–5). These results also highlight a few counties that have elevated *MLE* throughout the last 10–15 years (Fairfield, Nassau, New Haven, and Westchester), whereas others are still free of high *MLE*, as depicted in Fig 4 by the high number of red dots and black dots, respectively. These counties are also notable in the persistent accounts of reported human cases (Fig 5). Interestingly, Suffolk County showed high numbers of human cases over the years, whereas the WNV *MLE* are more temporally variable, likely due to our exclusion of light traps, which represent the majority of the trapping effort in this county. The year 2012 exhibits the highest number of counties with high *MLE* (10 of 14 counties with $MLE > 5$ infected per 1000). The model predicted 9 of those counties correctly, but underestimated risk in one county and overestimated it in 3 others (Fig 3). The years 2011 and 2013, before and after the peak year (2012) showed fewer counties with high *MLE*. The model generally reproduces region-wide variations, an indication that climatic conditions play a role in the WNV *MLE*.

The accuracy of predictions was variable for individual counties or years (Figs 4 and 5). The models retained both climatic and non-climatic variables in the final predictive models (Tables 4 and 5). In particular, *MLE* increased with increasing mean minimum temperature for July, August and September (Table 4, Fig 6). At the county scale, *MLE* showed a non-linear relationship to soil moisture in April, May and June. Years with low soil moisture were always at risk

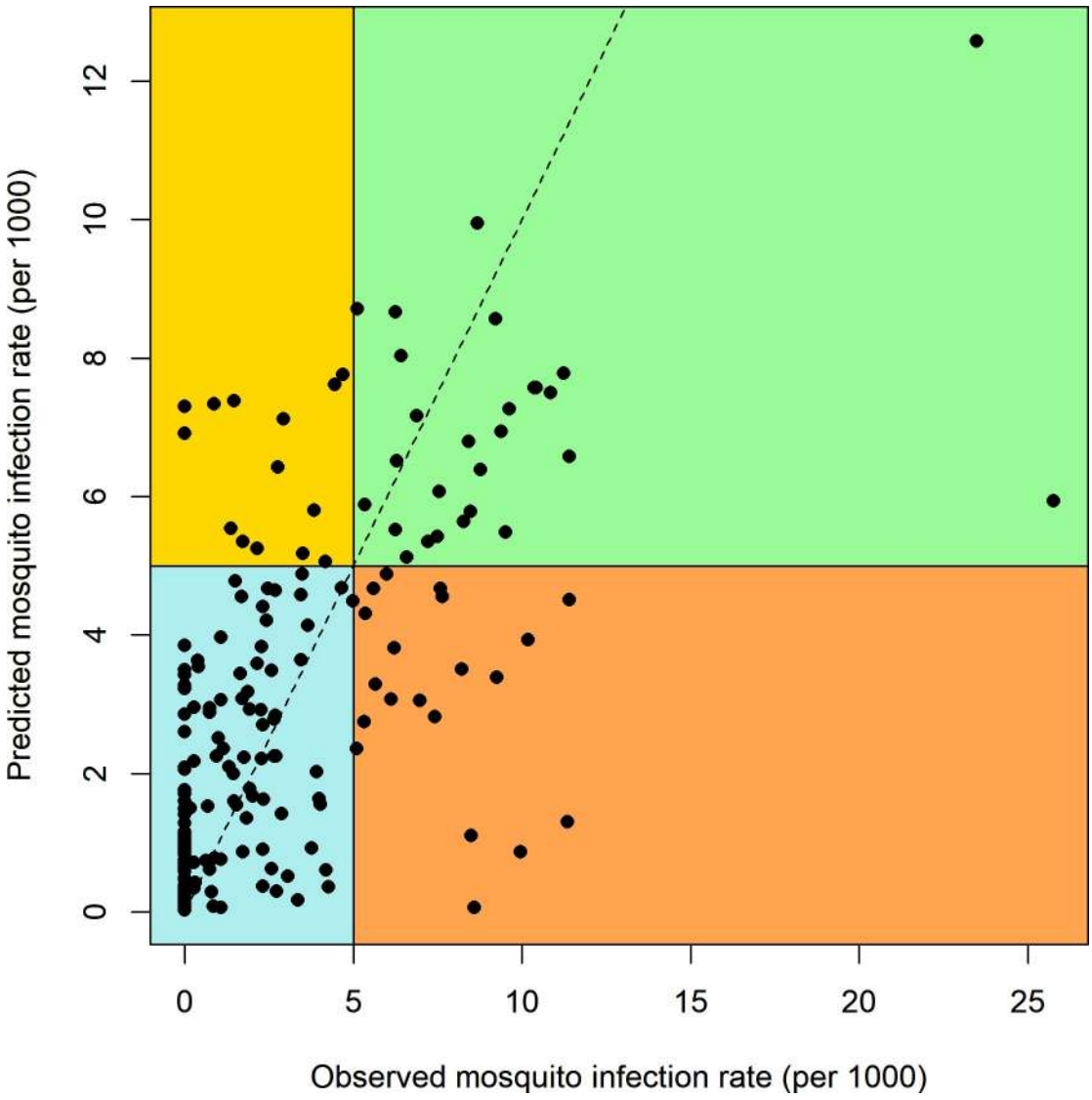

Fig 1. Observed mosquito infection rate (*MLE*) vs. predicted *MLE* from the WNV model using the entire data set. Background colors correspond to a classification of model predictions based on *MLE* of 5 [22]. Green corresponds to a correct prediction of high WNV *MLE* (27 records, 12.4%), blue corresponds to a correct prediction of low WNV *MLE* (157 records, 72.0%). Yellow corresponds to an error where the model predicts *MLE* to be high, but it is not (14 records, 6.4%), whereas orange corresponds to an error where the model predicts *MLE* to be low, but *MLE* was high (20 records, 9.2%). Future models should aim to improve the model's sensitivity (0.57), although the specificity (0.92) is also of concern. Note that some predictions can be quite accurate, and still result in misclassification if they are near the classification threshold.

https://doi.org/10.1371/journal.pone.0217854.g001

of high WNV, years with normal soil moisture corresponded to a risk of WNV when mean minimum temperatures were high, and years with above-normal soil moisture were associated with a slight increase in WNV risk relative to years with normal soil moisture and cool mean minimum temperatures (Fig 6). High mean minimum temperature in January, February and March was also associated with higher WNV rates, as were droughts in July, August and September (Fig 7). At the scale of individual traps, mosquito abundance and maximum observed temperature from April to June were among the most important variables (Table 4, Fig 8). American Robin abundance was also predictive of mosquito infection rates. Similar to the

Table 2. Model fit results for the calculated mosquito infection rates (per 1000). Climate indicates whether climate variables were included, *N* indicates sample size, while WNV+ N indicates the number of samples estimated to have WNV present. RMSE, Median RMSE, Max Error, Scaled RMSE, R^2, r_p, and r_s are defined in *Methods*: *model fit statistics*.

Scale	Climate	*N*	WNV+ *N*	RMSE	Median RMSE	Scaled RMSE	Max Error	R^2	r_s	r_p
County	YES	218	132	2.8	2.3	1.04	19.8	0.45	0.69	0.68
County	NO	218	132	3.3	2.7	1.21	23.3	0.26	0.67	0.51
Trap	YES	3156	955	8.2	7.7	2.34	87.4	0.16	0.45	0.40
Trap subset	YES	2596	395	1.2	1.0	2.13	15.4	0.53	0.59	0.73
Trap subset	NO	2596	395	1.4	1.2	2.49	10.9	0.36	0.57	0.61

https://doi.org/10.1371/journal.pone.0217854.t002

mosquitoes, human cases also show an increase with mean minimum temperature for July, August and September, and with total human population in the county (Fig 9).

Model fit with and without climate variables

Models without climate variables explained 17–19% less of the total variance in *MLE* than models with climate variables (Table 2) and removal of climate variables resulted in a poorer fit (Table 2). Although the mean errors were smaller with climate variables included, the maximum error observed was sometimes greater (i.e., when evaluated at the trap-subset scale). Removal of climate variables led to changes in which non-climatic variables were included in the models (Table 4). Models that include climate variables explained 7–12% more of the total variance for the number of human cases (Table 3). Again, removal of climate variables led to the inclusion of additional non-climatic variables into the model (Table 5). For both *MLE* and human cases, Pearson correlation coefficients were higher with climate variables included (Δr_p 0.12–0.17 for *MLE*, 0.04–0.07 for human cases, Tables 2 and 3), whereas Spearman correlations were often similar with and without climate variables (Δr_s 0.02 for *MLE*, -0.06–0.04 for human cases, Tables 2 and 3). This suggests that climate variables improve the estimation of the magnitude of WNV outbreaks across years; in contrast non-climatic variables may determine the baseline risk for a given location.

Model fit by scale

The trap-scale subset had the lowest RMSE for *MLE*, whereas the trap-scale had the highest RMSE (Table 2). The county-scale had the lowest RMSE when scaled by the mean infection rate. The percent of variation in *MLE* explained relative to the mean (R^2) was also greatest at the trap-scale subset (Table 2).

Description of key variables

Mean minimum temperatures (3rd quarter) were highly temporally correlated across counties (Fig 10A), although counties differed in their mean temperature. Soil moisture anomalies (2nd quarter) were also correlated, but showed some large, county-specific deviations (Fig 10B). Human cases often, but not always, tracked mosquito infection rates (Fig 10C and 10D). Total population was also identified as important by many of the models (Table 4), and the population distribution is presented in Fig 11.

Discussion

We found that climate variables improved WNV model fit metrics at both coarse and fine scales (with a few minor exceptions, see Tables 2 and 3). Climate variables were especially

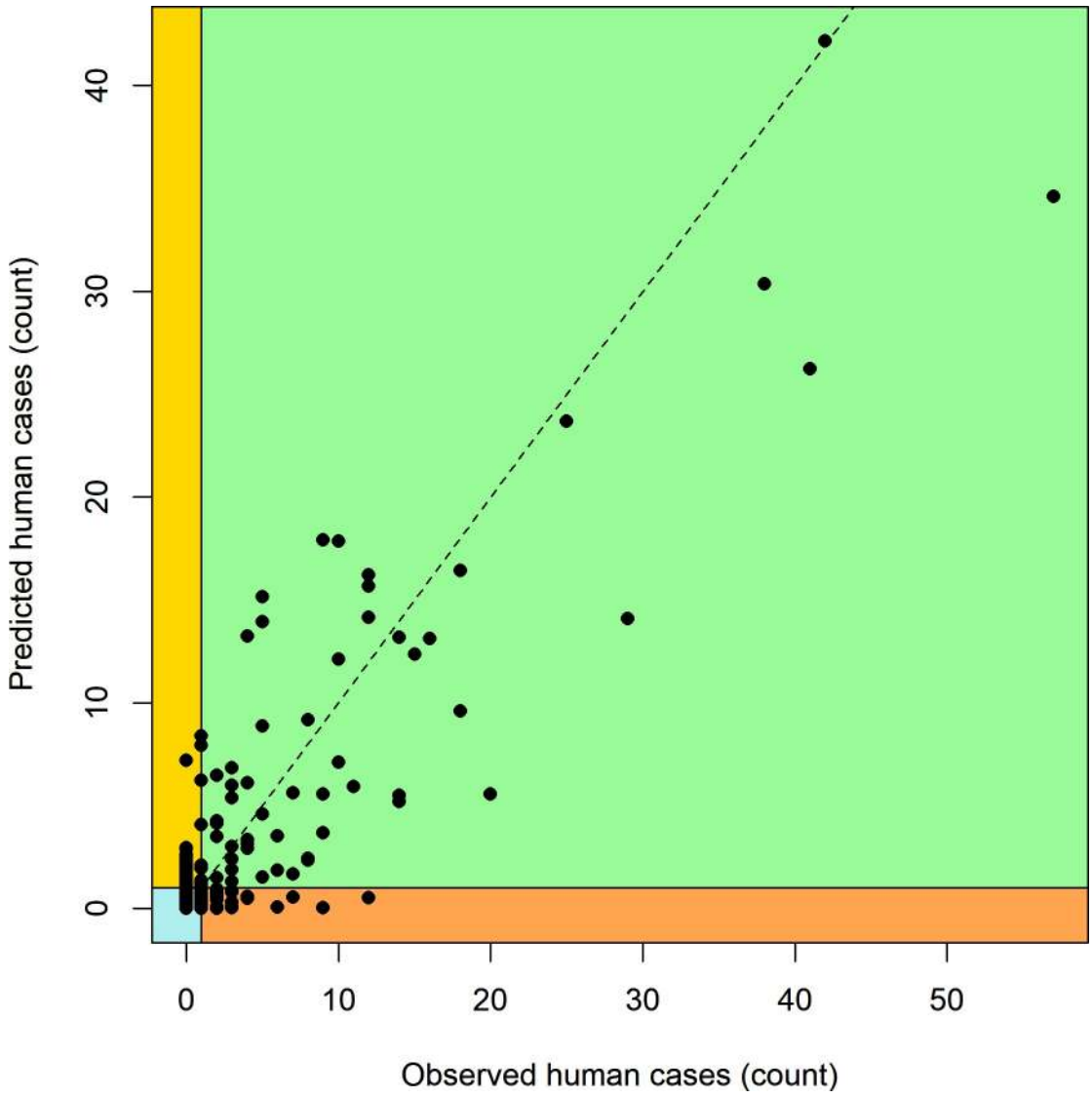

Fig 2. Observed number of human cases of WNV across all of New York and Connecticut vs. predicted number of human cases of WNV from the model using the entire data set. Background colors correspond to a classification of model predictions based on a threshold of 1 human case. Green corresponds to a correct prediction of one or more human cases (65 records, 7.4%), blue corresponds to a correct prediction of no human cases (704 records, 79.8%). Yellow corresponds to an error where the model predicts at least one human case, but none were observed (38 records, 4.3%), whereas orange corresponds to an error where the model predicts no human cases, but at least one was observed (75 records, 8.5%). Sensitivity (0.46) and specificity (0.95) were similar to the estimates for county-scale mosquito infection rates.

https://doi.org/10.1371/journal.pone.0217854.g002

important for predicting human West Nile cases across all counties ($\Delta R^2 = 0.12$), and mosquito *MLE* at the county- and trap-scales ($\Delta R^2 = 0.19$, $\Delta R^2 = 0.17$, respectively). We found evidence that some of the climate effects on WNV were an indirect result of climatic effects on mosquito populations (see e.g., [112]). When climate data were omitted, *MLE* became more important in predicting human cases and the mosquito abundance index became more important for explaining *MLE* at the trap scale (Table 5).

Within the climate predictor set, the mean minimum temperature in July, August and September was frequently included as a predictor in the best models. Human population was also consistently important in predicting both *MLE* and number of human cases (Table 5),

Table 3. Model fits for the human data at the county-scale. The All Counties analysis was based on 882 county × year records, while the subset contained 206 county × year records for which surveillance data were available. RMSE, Max Error, Median RMSE, Scaled RMSE, R^2, r_p, and r_s are defined in *Methods*: *model fit statistics*.

Scale	Climate	RMSE	Median RMSE	Scaled RMSE	Max Error	R^2	r_s	r_p
All Counties	YES	2.0	1.6	2.45	30.2	0.72	0.39	0.86
All Counties	NO	2.5	1.7	2.93	37.6	0.60	0.45	0.79
Subset	YES	3.7	1.7	1.80	42.3	0.52	0.70	0.72
Subset	NO	4.0	2.1	1.94	44.1	0.45	0.66	0.68
Subset -S[1]	YES	3.9	1.7	1.88	43.1	0.48	0.70	0.69

[1] Without surveillance variables

https://doi.org/10.1371/journal.pone.0217854.t003

consistent with the urban nature of *Cx*. *pipiens* and results of prior research (e.g., [39]. When climate variables were excluded, the importance of mosquito infection rate for predicting human cases greatly increased. This further supports a major role of climate mediating mosquito infection rates.

Our results are broadly consistent with nine previous studies that have included climate data and NY or CT in their scope (Table 1). To some degree, this was due to analyzing data that was also incorporated in the prior studies. For the analyses across the United States, our human case data represent a subset of their overall data set. In contrast, the data used in several more localized studies in CT, Suffolk County, and Erie County represent subsets of our larger data set. To our knowledge, the mosquito surveillance data set used here represents the largest data set applied to NY and CT at the county scale or below. Our study also differed from most previous analyses in this region (except [55]) in our use of machine learning techniques.

We found that the RMSE estimates produced by the models were highly scale-dependent. As the spatial scale changed, the estimated infection rates changed. For the trap-scale subset, this is due in part to dropping records with high uncertainty in the mean estimate. This disproportionately, but not exclusively, affected sites where WNV was detected. Second, the difference in estimated infection rates could be due to aggregating samples in the presence of spatial heterogeneity and unequal sampling [33]. This scale-dependence of fit statistics is important to consider when comparing results across studies, and highlights the value of standardizing the RMSE by the mean infection rate. Otherwise, one might conclude unequivocally that the trap-scale subset results were more accurate based on the RMSE. Although in terms of absolute error this is true, the lower mean infection rate indicates that there was less potential for error in that model and the scaled RMSE indicates greater error relative to the mean value. More broadly, while our model RMSE of 2.8 at the county-scale is similar to the RMSE of 4.3 observed by Little et al. [22] for Suffolk County, we note that the results are not strictly comparable as their model was evaluated at a 13 × 13 km scale, in contrast to our results that were at the county-scale, and no scaled RMSE was reported for that study. For humans, the RMSE of 2.0 indicates that our model predictions were off by an average of ±2 human cases. This number must be interpreted with caution, though. We speculate that much of this error was due to a few years with exceptionally high numbers of cases (maximum error of 30.2), as the median RMSE was 1.6 and squared errors are especially sensitive to large deviations (Table 3, All Counties results).

Here, we created predictive models with a minimum number of variables (minimum predictive models). It is worth noting that our variable selection approach did not identify all relevant covariates [113,114]. For example, changing the random forest starting seed changed which variables were included in the final model (not shown). This is due to two issues 1) collinearity among predictor variables (e.g., Fig 9) and 2) the complexity of the system (e.g., Fig

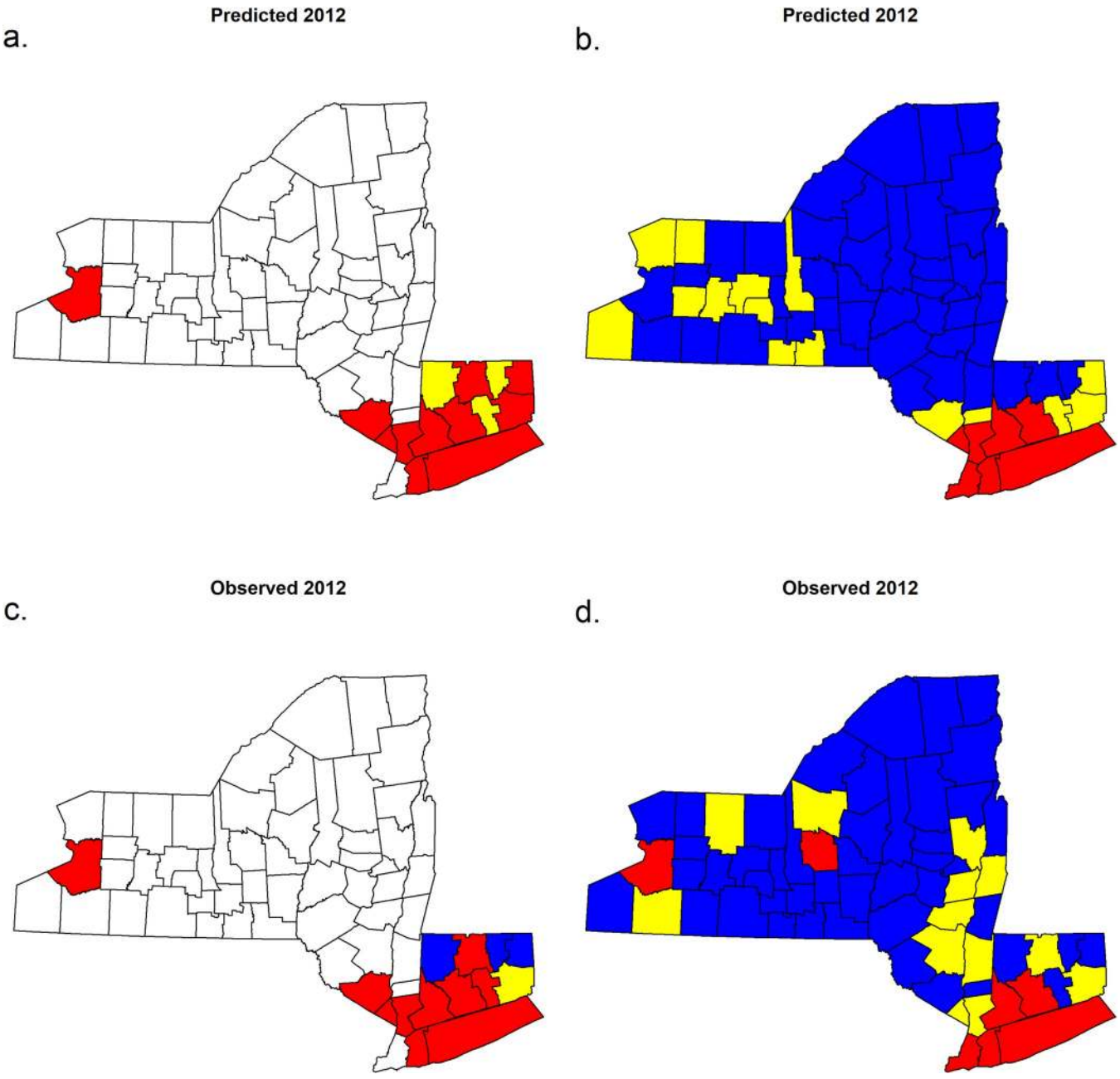

Fig 3. Predicted and observed WNV mosquito infection rates (*MLE*, a, c) and human cases (b, d) for 2012, a particularly widespread WNV year. *MLE* thresholds from Little et al. [22]: blue corresponds to $MLE < 1$ mosquito per 1000, yellow corresponds to *MLE* 1–5 per 1000, and red to $MLE > 5$ per 1000. White indicates excluded counties for which we did not have mosquito surveillance data. For human cases (b, d), blue indicates no human cases, yellow indicates 1–5 cases, and red indicates more than 5 cases.

https://doi.org/10.1371/journal.pone.0217854.g003

6). The high correlation among some of the variables (S7 File) may have obscured which variable(s) were the most important from a mechanistic standpoint. As Fig 9 shows, total human population and minimum 3rd quarter temperature are both potentially important for explaining observed infection rates, but the range of variation makes it difficult to separate the individual effects of these variables. This is further complicated by the complexity of the system. Fig 6 demonstrates the potential for interactions between variables, for example, where the

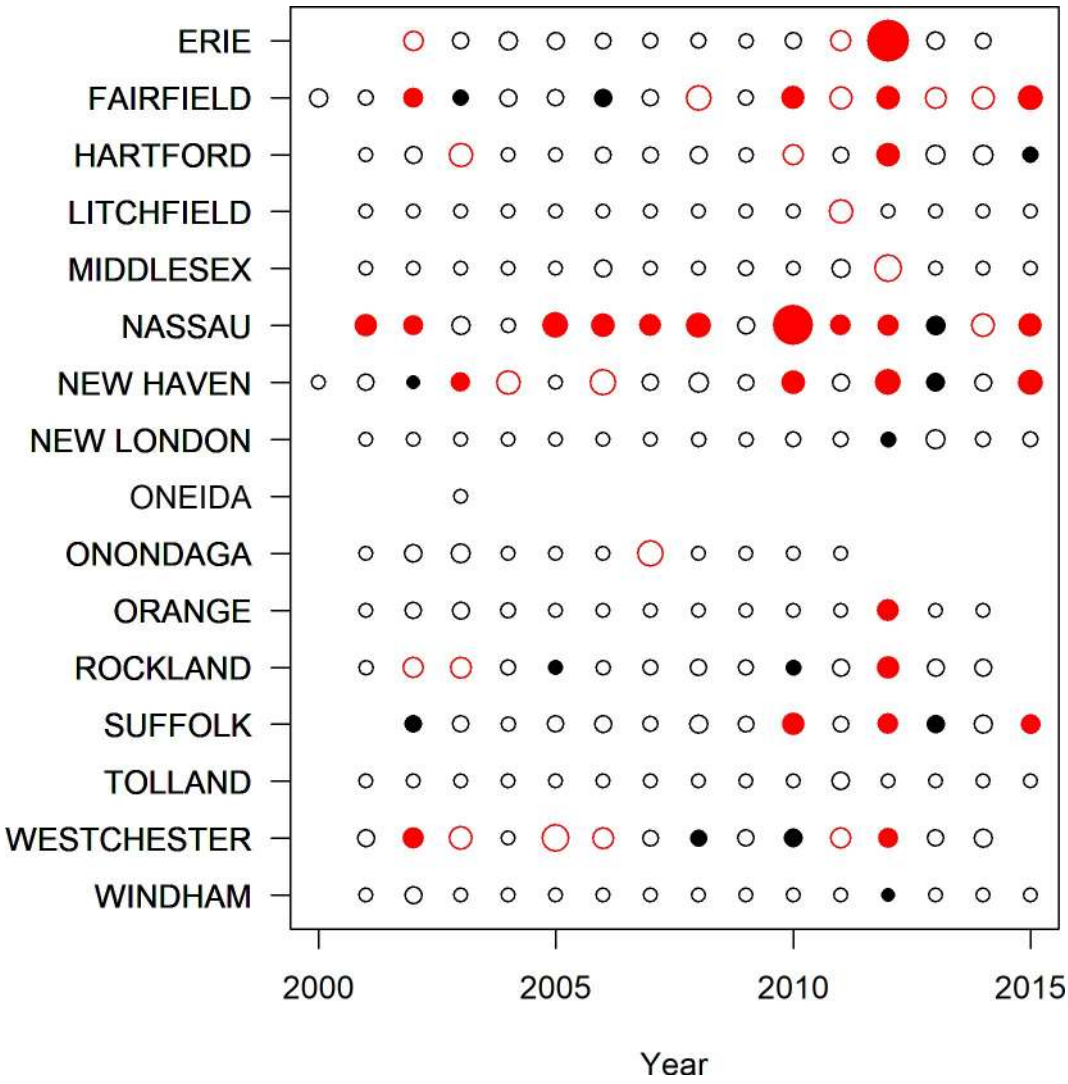

Fig 4. Predicted (unfilled $\leq$ 5, filled $>$ 5) and observed (black $\leq$ 5, red $>$ 5) infected mosquitoes per 1000 for each county and year for WNV. Missing points correspond to missing years for those counties. Point sizes are scaled relative to the observed infection rate.

https://doi.org/10.1371/journal.pone.0217854.g004

relationship with soil moisture depends on the temperature. Above a certain minimum temperature, there is no longer a strong relationship with soil moisture. There are many possible interactions among variables, and conclusively identifying them with small data sets may not be possible. We note that this issue is not unique to the random forest approach employed here. Little et al. [22], using linear methods, also found that multiple models with very different sets of variables had similar explanatory power for predicting WNV.

It is interesting to note, however, that the model predictions were very good when extreme climate conditions were encountered. Exceptionally warm average minimum temperatures in January–March and from July–September, were often associated with a much higher risk of WNV. Drought conditions reduced the minimum temperatures necessary for a WNV outbreak, or may have amplified the magnitude of the outbreak. In contrast, during the typical minimum temperature or during normal soil moisture, WNV risk was reduced, but more difficult to predict precisely, due to variation in *MLE* when climatic conditions were similar.

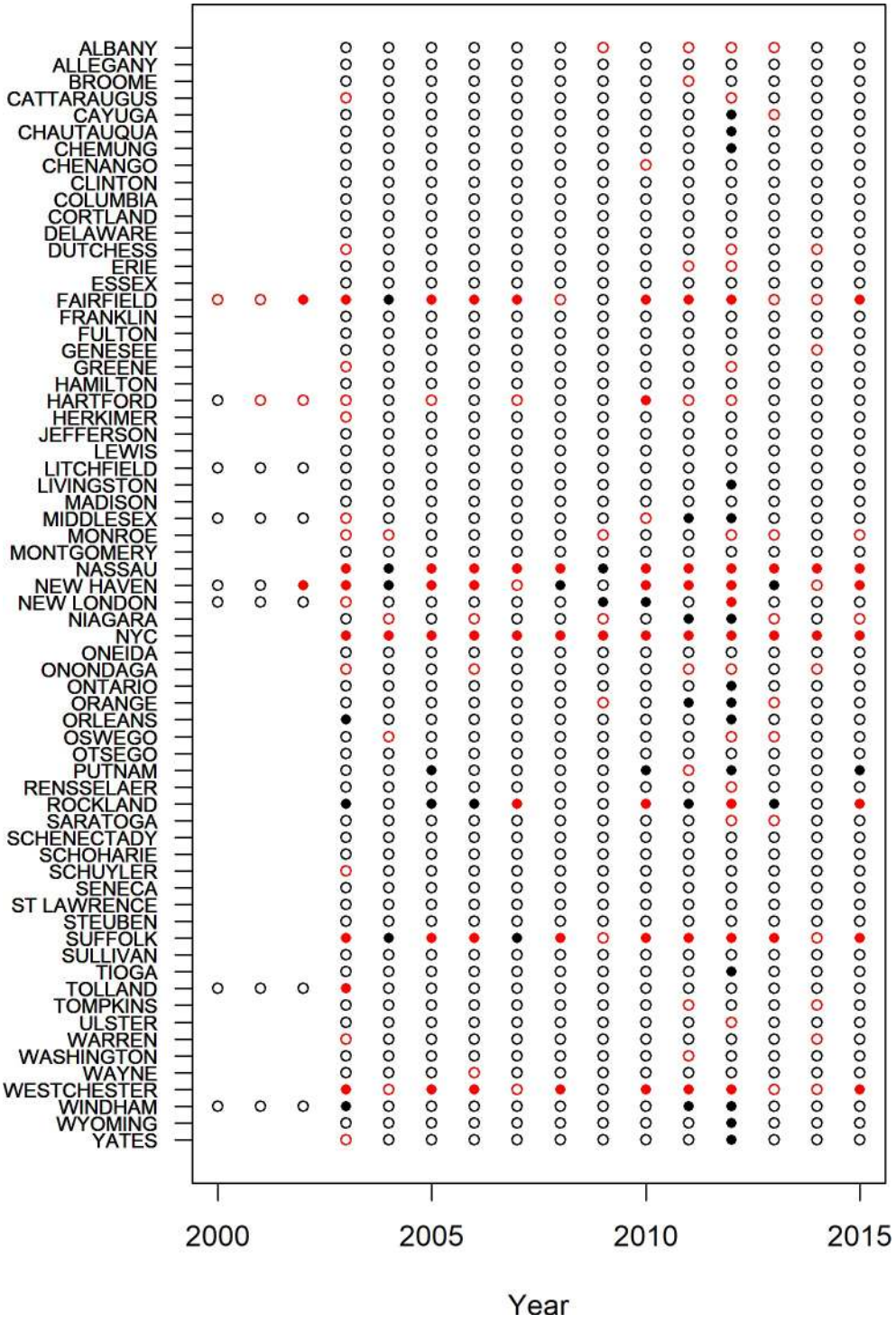

Fig 5. Predicted (open < 1, filled ≥ 1) and observed (black < 1, red ≥ 1) number of human WNV cases for each county and year. Data were not available for New York for 2000–2002, hence the missing points.

https://doi.org/10.1371/journal.pone.0217854.g005

Limitations of the present study

Future research could include additional covariates, such as multiple buffer distances at the trap-scale [46], the Normalized Vegetation Difference Index [43,46], socio-economic variables

Table 4. Climate variables identified as important by the random forest model when the model with all covariates was run, and when a model with only climate covariates was run (only C). Model results are presented for human cases in those counties where mosquito surveillance data were collected, and for mosquito infection rates (*MLE*) at both the county and trap scales. Values in the table indicate the amount of unique variation explained by the variable using variance partitioning, while a blank indicates that the variable was not included in the final predictive model.

Variables appearing in a final model	Human subset	Human subset only C	MLE county	MLE county only C	MLE Trap subset	MLE Trap subset only C
Mean minimum temperature (Jan–Mar)		<0.001	0.01	0.001		
Mean minimum temperature (Apr–Jun)						0.01
Mean minimum temperature (Jul–Sep)	0.004	0.03	0.01	0.02	0.01	0.01
Mean minimum temperature anomaly (Oct–Dec)						0.01[a]
Mean maximum temperature (Jan–Mar)			0.003	0.001		0.01
Mean maximum temperature (Jul–Sep)			0.001			
Mean maximum temperature anomaly (Jan–Mar)			0.002			
Minimum observed temperature (Jul–Sep)		0.01				
Minimum observed temperature (Oct–Dec)						0.01
Maximum observed temperature (Apr–Jun)	0.02	0.01			0.03	
Maximum observed temperature (Oct–Dec)				0.02[a]		
Maximum observed temperature anomaly (Apr–Jun)	0.02	0.01				
Daily temperature range (Jan–Mar)		0.003				
Daily temperature range (Jul–Sep)						0.01
Daily temperature range (Oct–Dec)				0.004[a]		
Daily temperature range anomaly (Jan–Mar)						0.02
Soil moisture anomaly (Apr–Jun)			0.03	0.04		
Soil moisture anomaly (Jul–Sep)			0.04	0.05		
Soil moisture anomaly (Oct–Dec)						0.01[a]
Growing degree days (Jul–Sep)		0.002				
Growing degree days anomaly (Apr–Jun)						0.01
Growing degree days anomaly (Oct–Dec)	0.01[a]	0.02[a]				0.01[a]

[a] We hypothesize that the contribution of this variable is related to the end of the mosquito season in October.

https://doi.org/10.1371/journal.pone.0217854.t004

Table 5. Non-climatic variables identified as important by the random forest model when the model with all covariates was run, and when a model without climate covariates was run (-C). Model results are presented for human cases in those counties where mosquito surveillance data were collected, and for mosquito infection rates (MLE) at both the county and trap scales. Values in the table indicate the amount of unique variation explained by the variable using variance partitioning.

Variables appearing in a final model	Human subset	Human subset -C	MLE County	MLE County -C	MLE Trap	MLE Trap -C
Mosquito infection rate	0.02	0.05	NA	NA	NA	NA
Mosquito abundance index					0.15	0.28[a]
Mosquito density index						0.06[a]
Trap bait type						
Total population	0.01	0.02	0.003	<0.001		
Population density		0.02		0.002	0.01	
Percent urban		0.02		<0.001	0.01	
Percent forest	0.002	0.03		<0.001		
Percent open		0.02		0.001		
Percent wetland				0.002		
American Robin Index		0.06		0.02	0.01	0.02[a]
American Crow Index		0.002		0.01		0.03[a]

[a] We note that the sum of the values in this column exceeds the total amount of variation explained by the model (0.36). This occurred because the model without one or more of these variables explained less variation than just using the mean value from the validation data set and therefore had a negative R^2 value as the baseline instead of zero (see *Coefficient of determination* section in methods for the method of calculating the R^2).

https://doi.org/10.1371/journal.pone.0217854.t005

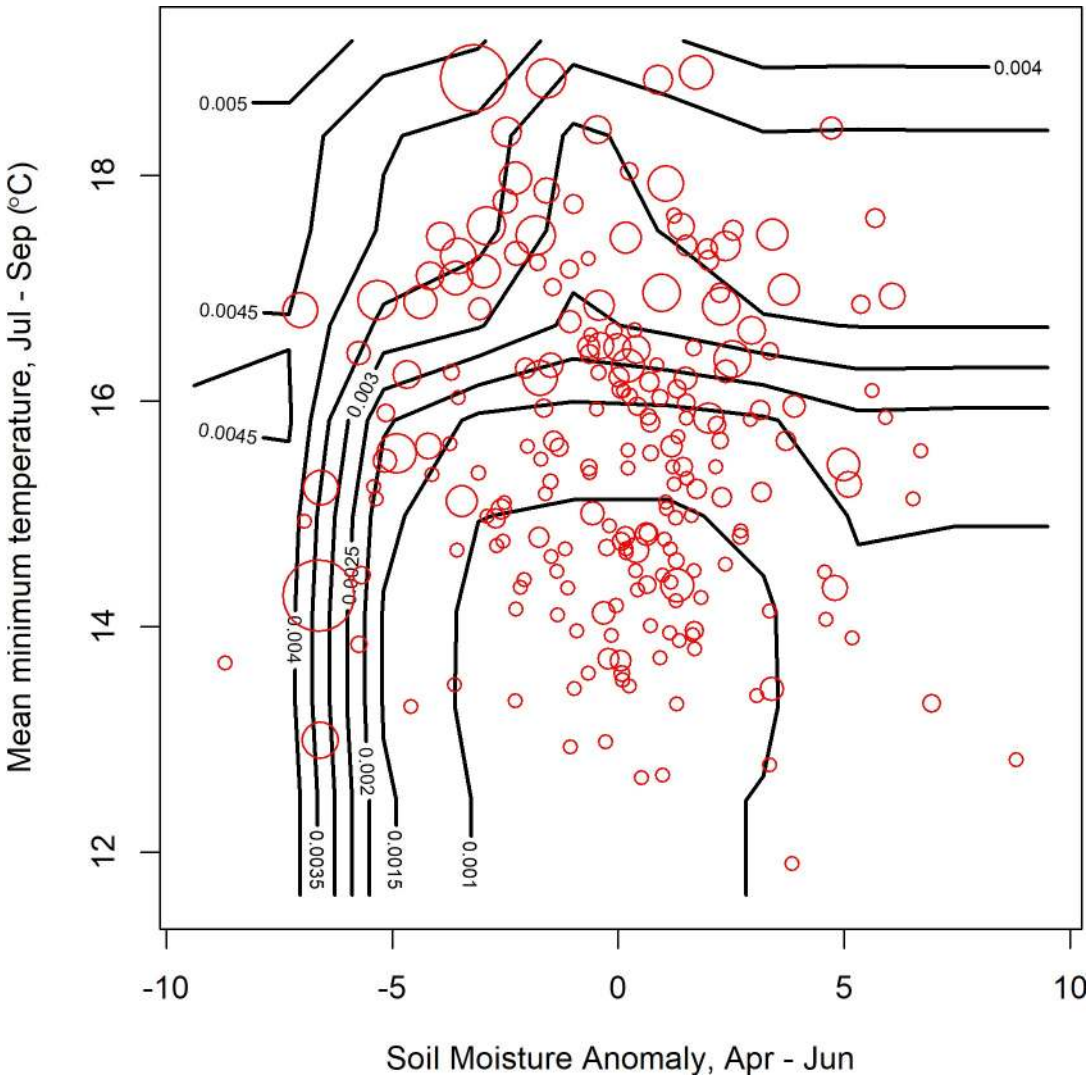

Fig 6. Predicted mosquito infection rates (*MLE*, contours) increase non-linearly with 2nd quarter soil moisture anomaly and 3rd quarter temperature. Cool years with normal soil moisture were associated with the lowest *MLE*. Warm years showed high *MLE* regardless of soil moisture and dry years often (but not always) had high *MLE*. Observations (red circles, size is proportional to *MLE*) broadly support these predictions. Contour lines correspond to predictions made for a regular grid of 100 points covering the range of both variables. Predictions were made for mean values for all other covariates (see Tables 4 and 5 for included variables, see S1 File for mean values), while observed values correspond to the exact variable combinations and therefore may not exactly correspond to the predictions. Observations are plotted as a general guide to identify major patterns and highlight particular exceptions.

https://doi.org/10.1371/journal.pone.0217854.g006

(especially age) [52,115–117], changes in host behavior (related to a shift in mosquito feeding preferences) [20], rates of host immunity, rate of human immunity [31], degree of *Cx. pipiens pipiens* × *Cx. pipiens molestus* hybridization due to changes in contact rates [74], mosquito control activities and lagged climate effects [28,29]. Of the omitted covariates, human age, mosquito control activities, and climate-lags may be the most critical. Age is a major factor in whether WNV becomes neuroinvasive [1,118], and the number of susceptible humans could be an important consideration [31]. However, one study found reduced WNV in areas with elderly populations due to those populations being located in areas that were less risky for WNV based on degree of urbanization [52]. Mosquito control efforts could contribute to a

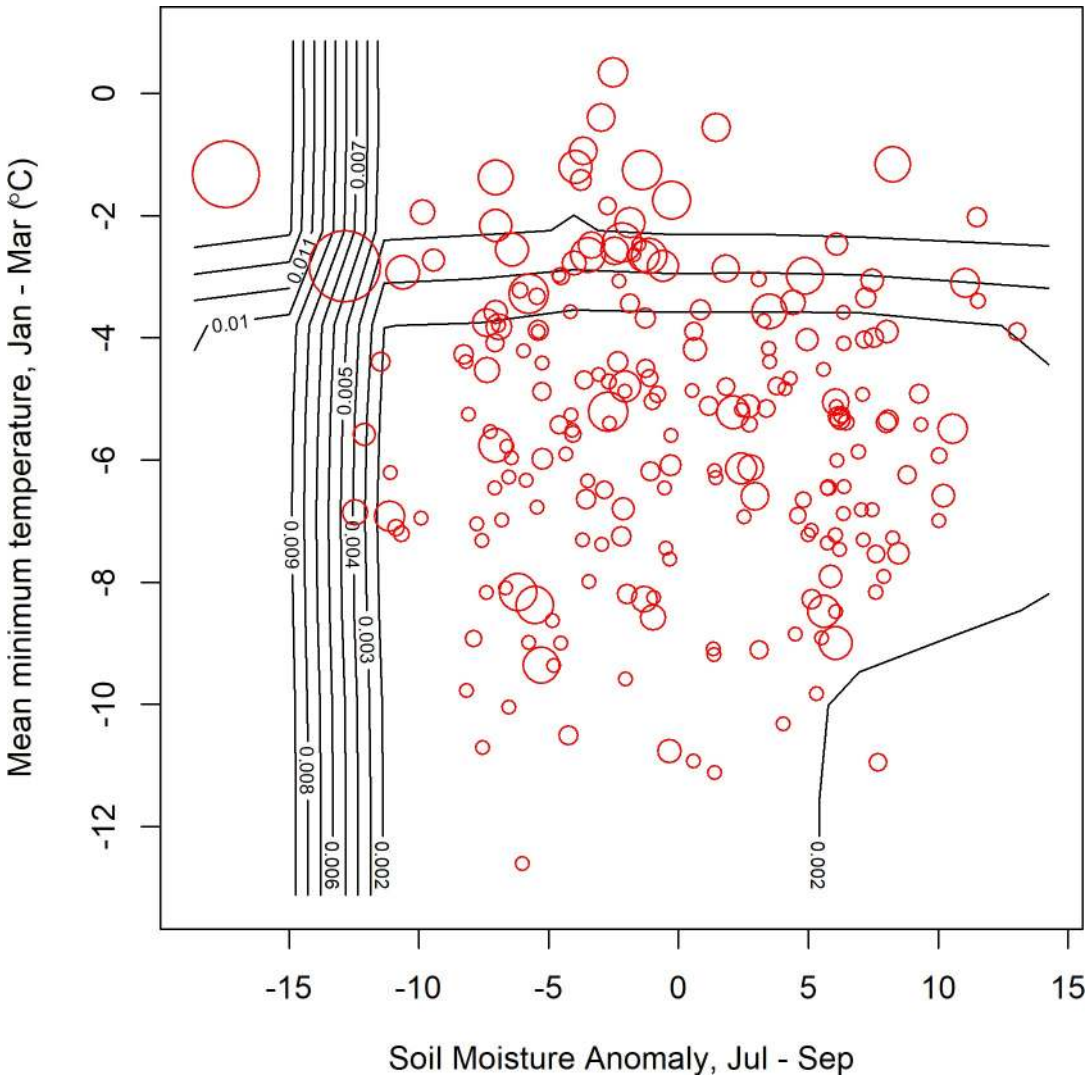

Fig 7. Warm winter temperatures and dry summers were associated with the highest risk of mosquito infection with WNV. Observations (red circles, size is proportional to infection rate) broadly support these predictions. Contour lines correspond to predictions made for a regular grid of 100 points covering the range of both variables. Predictions were made for mean values for all other covariates (see Tables 4 and 5 for included variables, see S1 File for mean values), while observed values correspond to the exact variable combinations and therefore may not exactly correspond to the predictions. Observations are plotted as a general guide to identify major patterns and highlight particular exceptions.

https://doi.org/10.1371/journal.pone.0217854.g007

mismatch between predictions and observations. If conditions are suitable for WNV, but mosquitoes have been controlled, the model may predict high WNV risk, but the actual risk may be low. Conversely, if mosquito control is included in the training data set, predictions for areas where control is absent but risk is otherwise high, could be low. These variables have been difficult to include: to our knowledge, one study included detailed information on the number of mosquito complaints and number of known larval sites [52], but none have included detailed spatial information of mosquito control activities. We suggest that such a data set would be highly beneficial. Our study, in contrast to others (e.g., [28,29]), did not consider lagged climate effects. In particular, prior-year precipitation has been found to influence WNV [29]. However, we note that several of the variables identified here would be available by

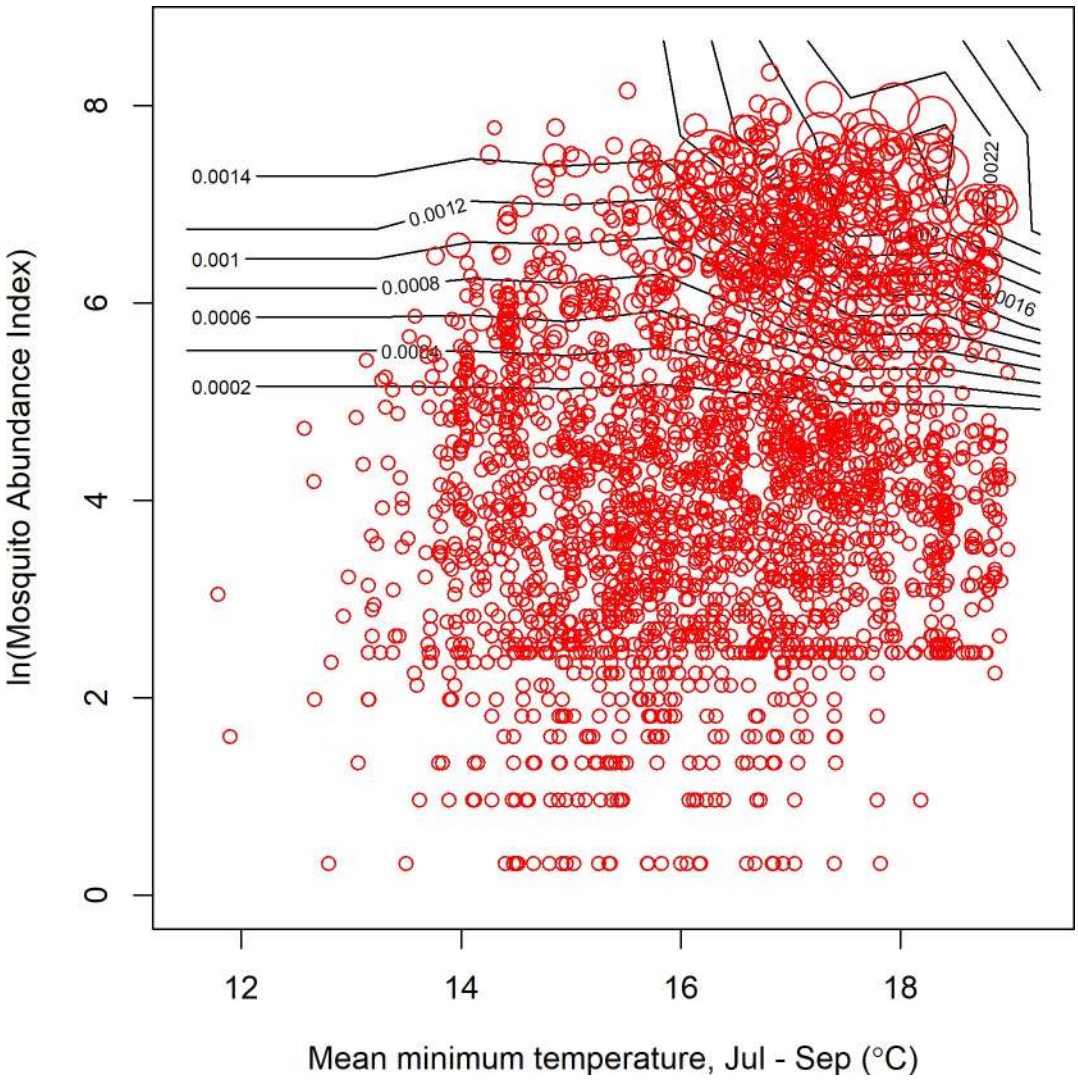

Fig 8. For individual trap sites, the risk of WNV increased with increasing mosquito abundance, especially when the mean minimum temperature in the 3rd quarter was high. Contour lines correspond to predictions from a regular grid of 100 points, (with values from other covariates fixed at a mean value). Observed infection rates (red circles, size is proportional to infection rate) are plotted for comparison, but note that they use exact parameter combinations and not the mean conditions used for making the predictions.

https://doi.org/10.1371/journal.pone.0217854.g008

early April or early July, and therefore could provide some predictive skill prior to the onset of human West Nile cases.

Statistically, the methods employed here could be further refined. Spatial and temporal autocorrelation may substantially influence model results [119,120]. We did not detect evidence of temporal or spatial autocorrelation based on a visual inspection of the model residuals [55]. It is possible that more refined models with respect to spatial or temporal autocorrelation could result in further improvements to the model fit statistics. However, we believe the process of evaluating model fit based on a validation data set indicates that our results are not a simple result of autocorrelation. Additionally, a prior study found no benefit to including a spatial autoregressive coefficient [25], but see [42].

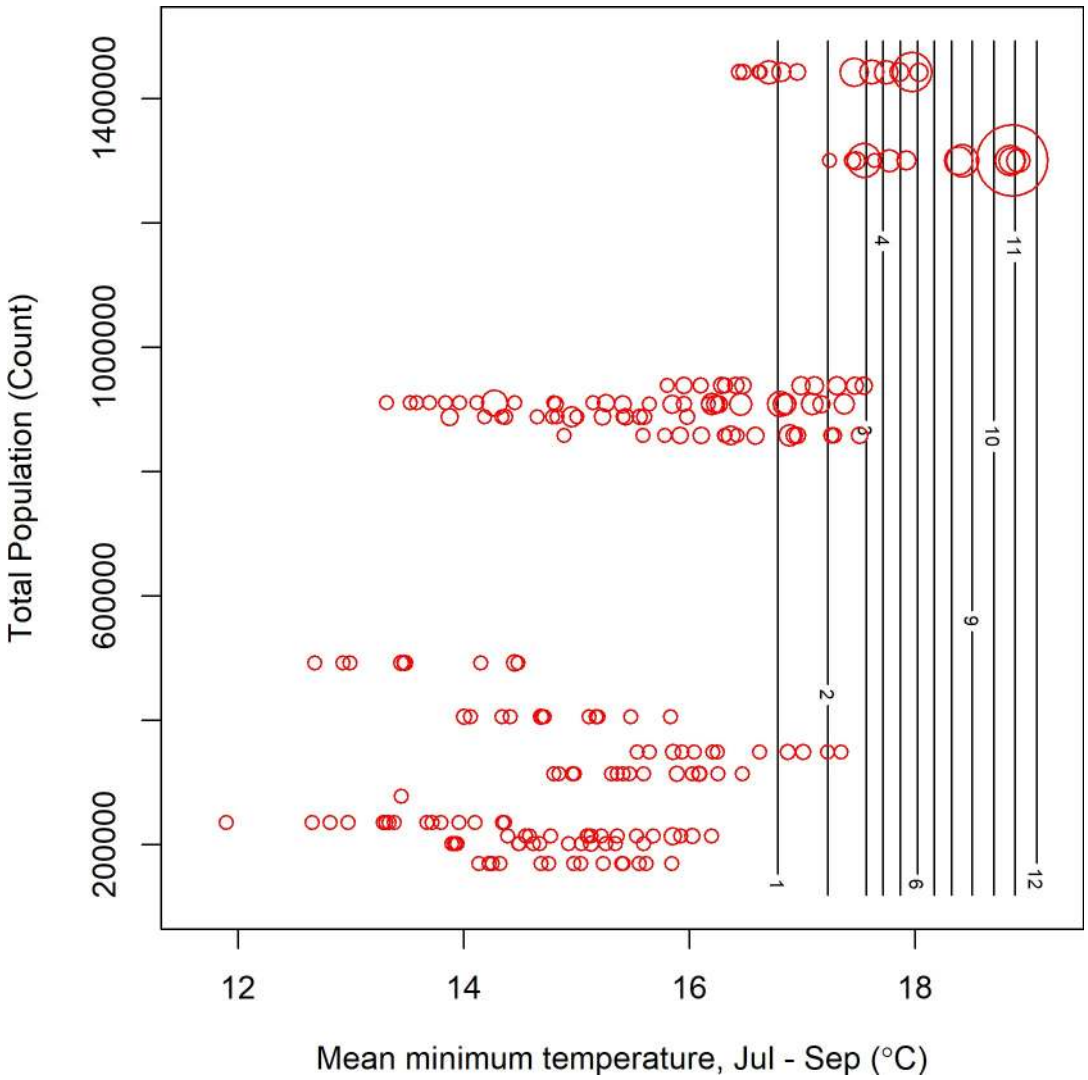

Fig 9. Risk of human cases of West Nile were highest for locations with high total populations, especially in years with a warm summer. Data correspond to the human subset analysis. Contour lines correspond to predictions from a regular grid of 100 points, (with values from other covariates fixed at a mean value). Observed infection rates (red circles, size is proportional to infection rate) are plotted for comparison, but note that they use exact parameter combinations and not the mean conditions used for making the predictions.

https://doi.org/10.1371/journal.pone.0217854.g009

Spearman correlations were not very different between the climate and non-climate models. One contributing factor may be the difficulty in obtaining unbiased results from rank order correlation statistics in the presence of zero-inflated data [121]. When WNV is absent, it creates a multi-way tie for the last rank. In contrast, the random forest model generated continuous estimates of WNV risk, making predicted ties unlikely. Consequently, a model could have a very low absolute error but still have a low Spearman correlation in the presence of zero-inflated data.

Some of the methodological decisions made in this study may also have influenced the lack of model fit. We restricted our analysis to gravid traps, and this decision under-sampled *Cx. salinarius*, as this species is trapped in greater numbers at light traps. In addition, other researchers have stated that gravid counts can be negatively affected when there are other

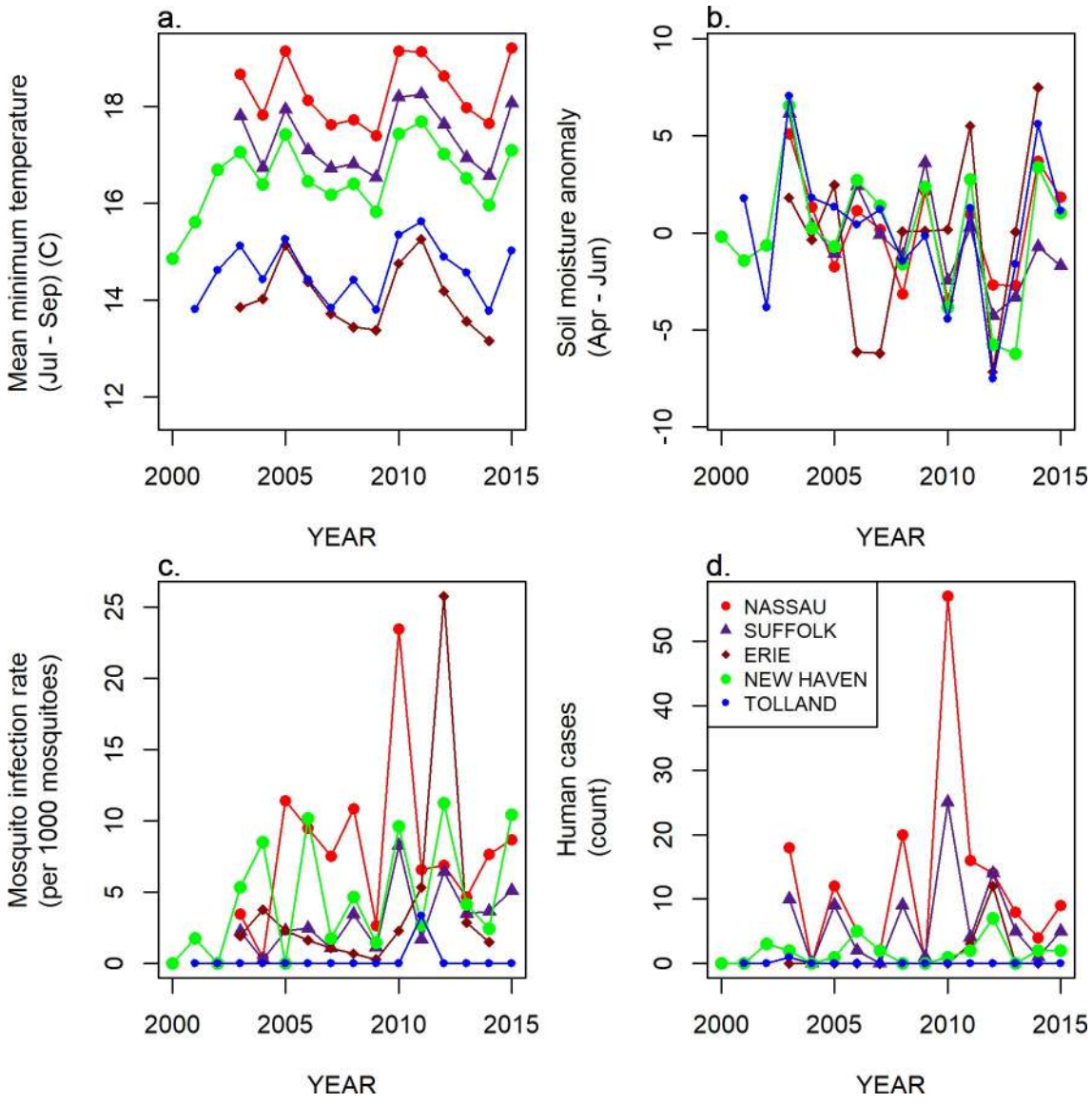

Fig 10. Mean minimum temperature (a), soil moisture anomaly (b), mosquito infection rate (c), and human case counts (d) by year for five example counties.

https://doi.org/10.1371/journal.pone.0217854.g010

sources of stagnant water [28], providing a possible bias towards fewer mosquitoes collected when there is higher precipitation. Trap success may also depend on the "pungency" of the trap water [28]. We considered only *Cx. pipiens*, *Cx. restuans*, and *Cx. salinarius* pooled together, as these three species were responsible for the majority of WNV positive pools in our data set, although 33 mosquito species in the Northeast [122] and at least 59 species worldwide have tested positive for WNV [64]. Importantly, the biology of *Cx. pipiens*, *Cx. restuans*, and *Cx. salinarius* differ, and pooling them may have increased the variation in our study. Methods have been developed to integrate multiple mosquito species into a single model [12], however this approach requires information on mosquito feeding preferences, which can vary spatially and temporally, even within a species [20,70]. It is possible that an analysis that spanned the entire mosquito community could improve the prediction of WNV in humans and mosquito

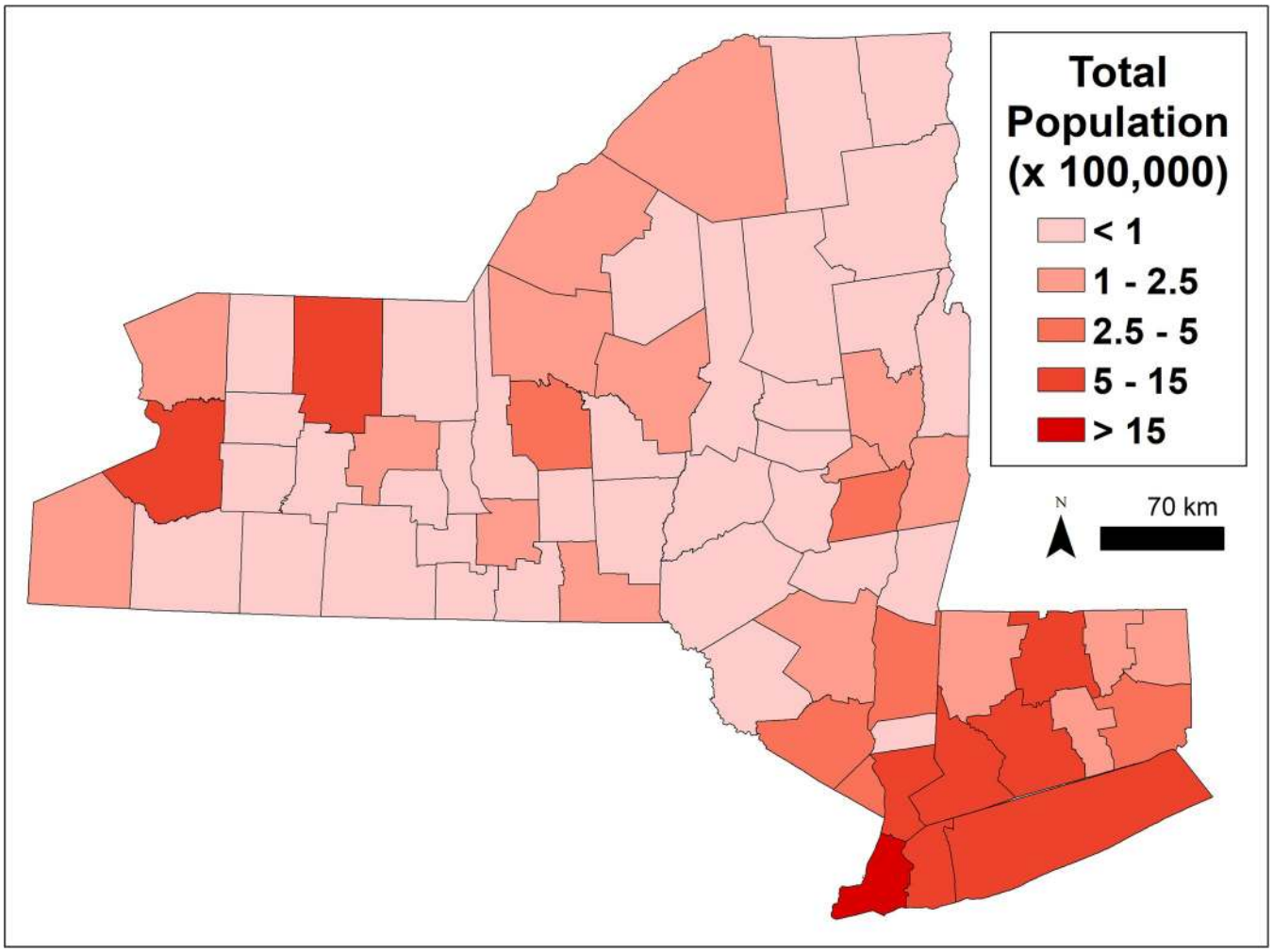

Fig 11. Total human population of the study region. Note that the five counties of New York City have been merged into a single entity. Data taken from the US Census [100,111].

https://doi.org/10.1371/journal.pone.0217854.g011

pools. An additional decision was to use the climate from the centroid of each county for the county-scale. Visual inspection suggested that our climate data were similar across counties (we compare the centroid to the county average in S6 File). The lack of standardization in the methods used to collect the mosquito data (S5 File) may have increased the variance associated with sampling error, and thereby contributed to the remaining unexplained variation. For example, trap sites in three CT counties were primarily urban/suburban, whereas those in the remaining five counties were primarily rural and sparsely populated. Sampling in NY was not independent of the presence of WNV, and this may have biased our results. The precise timing of sampling varied from county to county, and from year to year (S5 File). This could potentially bias the results, as more sampling outside the peak WNV season is expected to lead to lower overall annual *MLE*.

Future directions

Climate data from programs such as the Subseasonal to Seasonal (S2S) prediction project [123] and Subseasonal Experiment (SubX) provide novel opportunities for developing

predictive models for WNV prevalence. Aside from the inherent uncertainties in seasonal climatic predictions, the success of seasonal predictive tools targeting infectious diseases such as WNV will ultimately depend on the robustness of the connecting links between climate and the targeted biological-epidemiological system.

Conclusion

The WNV model developed here demonstrated predictive skill at multiple spatial scales for mosquito infection rates and human West Nile cases (see R^2 values, Tables 2 and 3). Including climate data improved model predictions substantially, as evidenced by the ability to explain a higher fraction of the total variance in the validation data. The applied random forest model appears to provide a valuable and highly adaptable statistical tool for the prediction of infectious spatial and temporal disease. However, it must be emphasized that more research is needed to improve the understanding of the mechanistic processes.

One of the remaining challenges is to deploy the model predictions in decision-making processes. Prediction errors could lead to costly action (in terms of time and money) when no increased WNV risk is present, or costly inaction (in terms of human and ecological health) when an increased WNV risk is present but not predicted by the model. Model errors, as demonstrated by the maximum errors, were sometimes substantial, although this may in part reflect the uncertainty in the estimated mean infection rates for the sampling units. Therefore, improved models likely require further refinement to be useful in an operational context. However, the model has heuristic value in helping to understand the dynamics of WNV and may be useful when extreme climatic conditions are present and risks of WNV are straightforward.

Supporting information

S1 File. Descriptive Statistics (.zip containing .csv files): Mean, standard deviation, median, minimum, maximum, the median, minimum, and maximum range observed within years, the median, minimum and maximum range observed across different years for variables included in the model. (1_1 for county_annual_mosquito, 1_2 for county_annual_human, 1_3 for point_annual_mosquito).
(ZIP)

S2 File. Data Dictionary for variable names.
(CSV)

S3 File. Data used to run the models at the county scales (.zip containing .csv files, see Data dictionary for variable names). S3_1 for county_annual_mosqutioes, S3_2 for county_annual_human, and S3_3 for county_annual_human_subset.
(ZIP)

S4 File. A comparison of the model results for incidence and total cases.
(DOCX)

S5 File. Mosquito sampling methods.
(DOCX)

S6 File. Comparison of climate data based on centroid and based on an average.
(DOCX)

S7 File. Correlations by scale (.zip containing .csv files): Bivariate correlations (Pearson) between all the dependent and independent variables used in this study for each spatial

scale (S7_1 for county_annual_mosquitoes, S7_2 for county_annual_human, S7_3 for county_annual_human_subset, S7_4 for trap_annual_mosquitoes for all trap sites, S7_5 for trap_annual_mosquitoes for the high-quality subset of trap sites).
(ZIP)

Acknowledgments

Disclaimer: This publication was supported by cooperative agreement 1U01CK000509-01, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

We thank Á. Muñoz and P. Williams for constructive discussion, C. Thorncroft for assisting with conceptualization and obtaining funding support, A. Rowe and J.L. White, and B. Laniewicz for access to NY mosquito data, K. D'Amico for assisting with data cleaning, and L. Harrington, E. Mader, and the Northeast Regional Center for Excellence in Vector-Borne Disease. We thank state and local agencies that do mosquito collection as part of arbovirus surveillance. We thank J. Dias and the University at Albany Vice President for Research Office for initial funding in support of this collaboration.

Author Contributions

Conceptualization: Oliver Elison Timm, P. Bryon Backenson, Kathleen A. McDonough, Mathias Vuille, Jan E. Conn, Laura D. Kramer.

Data curation: Alexander C. Keyel, Oliver Elison Timm, P. Bryon Backenson, Catharine Prussing, Sarah Quinones, Philip M. Armstrong, Theodore G. Andreadis.

Formal analysis: Alexander C. Keyel.

Funding acquisition: Oliver Elison Timm, P. Bryon Backenson, Kathleen A. McDonough, Mathias Vuille, Jan E. Conn, Laura D. Kramer.

Investigation: Alexander C. Keyel.

Methodology: Alexander C. Keyel, Oliver Elison Timm, Catharine Prussing, Sarah Quinones.

Project administration: Oliver Elison Timm.

Resources: Oliver Elison Timm.

Software: Oliver Elison Timm.

Supervision: Oliver Elison Timm, P. Bryon Backenson, Kathleen A. McDonough, Mathias Vuille, Jan E. Conn, Laura D. Kramer.

Validation: Alexander C. Keyel.

Visualization: Alexander C. Keyel, Oliver Elison Timm.

Writing – original draft: Alexander C. Keyel.

Writing – review & editing: Oliver Elison Timm, P. Bryon Backenson, Catharine Prussing, Sarah Quinones, Kathleen A. McDonough, Mathias Vuille, Jan E. Conn, Philip M. Armstrong, Theodore G. Andreadis, Laura D. Kramer.

References

1. CDC. Final Cumulative Maps & Data for 1999–2016 [Internet]. Center for Disease Control; 2017. Available: https://www.cdc.gov/westnile/statsmaps/cumMapsData.html

2. CDC. Species of dead birds in which West Nile virus has been detected, United States, 1999–2016 [Internet]. Center for Disease Control; 2017. Available: https://www.cdc.gov/westnile/resources/pdfs/BirdSpecies1999-2016.pdf
3. LaDeau SL, Kilpatrick AM, Marra PP. West Nile virus emergence and large-scale declines of North American bird populations. Nature. 2007; 447: 710. https://doi.org/10.1038/nature05829 PMID: 17507930
4. George TL, Harrigan RJ, LaManna JA, DeSante DF, Saracco JF, Smith TB. Persistent impacts of West Nile virus on North American bird populations. Proc Natl Acad Sci. 2015; 112: 14290. https://doi.org/10.1073/pnas.1507747112 PMID: 26578774
5. BirdLife International. Pica nutalli. The IUCN Red List of Threatened Secies 2016: e. T22705874A94039098 [Internet]. 2016. Available: http://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T22705874A94039098.en
6. Klenk K, Snow J, Morgan K, Bowen R, Stephens M, Foster F, et al. Alligators as West Nile virus amplifiers. Emerg Infect Dis. 2004; 10: 2150. https://doi.org/10.3201/eid1012.040264 PMID: 15663852
7. Root JJ, Oesterle PT, Nemeth NM, Klenk K, Gould DH, Mclean RG, et al. Experimental infection of fox squirrels (*Sciurus niger*) with West Nile virus. Am J Trop Med Hyg. 2006; 75: 697–701. PMID: 17038697
8. Schmidt JR, Mansoury HKE. Natural and experimental infection of Egyptian equines with West Nile virus. Ann Trop Med Parasitol. 1963; 57: 415–427. PMID: 14101930
9. Bowen RA, Nemeth NM. Experimental infections with West Nile virus. Curr Opin Infect Dis. 2007; 20: 293–297. https://doi.org/10.1097/QCO.0b013e32816b5cad PMID: 17471040
10. Kostyukov M, Alekseev A, Bulychev V, Gordeeva Z. Experimental infection of *Culex pipiens* mosquitoes with West Nile virus by feeding on infected *Rana ridibunda* frogs and its subsequent transmission (in Russian). Med Parasitol Mosc. 1986; 6: 76–78.
11. Mongoh MN, Hearne R, Dyer N, Khaitsa M. The economic impact of West Nile virus infection in horses in the North Dakota equine industry in 2002. Trop Anim Health Prod. 2008; 40: 69–76. PMID: 18551781
12. Kilpatrick AM, Kramer LD, Campbell SR, Alleyne EO, Dobson AP, Daszak P. West Nile virus risk assessment and the bridge vector paradigm. Emerg Infect Dis. 2005; 11: 425. https://doi.org/10.3201/eid1103.040364 PMID: 15757558
13. Work TH, Hurlbut HS, Taylor R. Indigenous Wild Birds of the Nile Delta as Potential West Nile Virus Circulating Reservoirs1. Am J Trop Med Hyg. 1955; 4: 872–888. PMID: 13259011
14. Komar N, Langevin S, Hinten S, Nemeth N, Edwards E, Hettler D, et al. Experimental infection of North American birds with the New York 1999 strain of West Nile virus. Emerg Infect Dis. 2003; 9: 311. https://doi.org/10.3201/eid0903.020628 PMID: 12643825
15. Kilpatrick AM, Daszak P, Jones MJ, Marra PP, Kramer LD. Host heterogeneity dominates West Nile virus transmission. Proc R Soc Lond B Biol Sci. 2006; 273: 2327–2333.
16. Deichmeister JM, Telang A. Abundance of West Nile virus mosquito vectors in relation to climate and landscape variables. J Vector Ecol. 2011; 36: 75–85. https://doi.org/10.1111/j.1948-7134.2011.00143.x PMID: 21635644
17. Reisen WK, Fang Y, Martinez VM. Effects of Temperature on the Transmission of West Nile Virus by *Culex tarsalis* (Diptera: Culicidae). J Med Entomol. 2006; 43: 309–317. https://doi.org/10.1603/0022-2585(2006)043[0309:EOTOTT]2.0.CO;2 PMID: 16619616
18. Dohm DJ, O'Guinn ML, Turell MJ. Effect of environmental temperature on the ability of *Culex pipiens* (Diptera: Culicidae) to transmit West Nile virus. J Med Entomol. 2002; 39: 221–225. https://doi.org/10.1603/0022-2585-39.1.221 PMID: 11931261
19. Richards SL, Mores CN, Lord CC, Tabachnick WJ. Impact of extrinsic incubation temperature and virus exposure on vector competence of *Culex pipiens quinquefasciatus* Say (Diptera: Culicidae) for West Nile virus. Vector-Borne Zoonotic Dis. 2007; 7: 629–636. https://doi.org/10.1089/vbz.2007.0101 PMID: 18021028
20. Kilpatrick AM, Kramer LD, Jones MJ, Marra PP, Daszak P. West Nile virus epidemics in North America are driven by shifts in mosquito feeding behavior. PLoS Biol. 2006; 4: e82. https://doi.org/10.1371/journal.pbio.0040082 PMID: 16494532
21. Vanderhoff N, Pyle P, Patten MA, Sallabanks R, James FC. American Robin (*Turdus migratorius*), version 2.0. Birds of North America. Ithaca, NY: Cornell Lab of Ornithology; 2016. Available: https://doi.org/10.2173/bna.462
22. Little E, Campbell SR, Shaman J. Development and validation of a climate-based ensemble prediction model for West Nile Virus infection rates in Culex mosquitoes, Suffolk County, New York. Parasit Vectors. 2016; 9: 443. https://doi.org/10.1186/s13071-016-1720-1 PMID: 27507279

23. Manore CA, Davis JK, Christofferson RC, Wesson DM, Hyman JM, Mores CN. Towards an early warning system for forecasting human West Nile virus incidence. PLoS Curr. 2014; 6.
24. Hahn MB, Monaghan AJ, Hayden MH, Eisen RJ, Delorey MJ, Lindsey NP, et al. Meteorological conditions associated with increased incidence of West Nile virus disease in the United States, 2004–2012. Am J Trop Med Hyg. 2015; 92: 1013–1022. https://doi.org/10.4269/ajtmh.14-0737 PMID: 25802435
25. Shaman J, Harding K, Campbell SR. Meteorological and hydrological influences on the spatial and temporal prevalence of West Nile virus in Culex mosquitoes, Suffolk County, New York. J Med Entomol. 2011; 48: 867–875. https://doi.org/10.1603/me10269 PMID: 21845947
26. Liu A, Lee V, Galusha D, Slade MD, Diuk-Wasser M, Andreadis T, et al. Risk factors for human infection with West Nile Virus in Connecticut: a multi-year analysis. Int J Health Geogr. 2009; 8: 67. https://doi.org/10.1186/1476-072X-8-67 PMID: 19943935
27. Walsh MG. The role of hydrogeography and climate in the landscape epidemiology of West Nile virus in New York State from 2000 to 2010. PLoS One. 2012; 7: e30620. https://doi.org/10.1371/journal.pone.0030620 PMID: 22328919
28. Trawinski P, Mackay D. Meteorologically conditioned time-series predictions of West Nile virus vector mosquitoes. Vector-Borne Zoonotic Dis. 2008; 8: 505–522. https://doi.org/10.1089/vbz.2007.0202 PMID: 18279008
29. Landesman WJ, Allan BF, Langerhans RB, Knight TM, Chase JM. Inter-annual associations between precipitation and human incidence of West Nile virus in the United States. Vector-Borne Zoonotic Dis. 2007; 7: 337–343. https://doi.org/10.1089/vbz.2006.0590 PMID: 17867908
30. Shaman J, Day JF, Stieglitz M. Drought-induced amplification and epidemic transmission of West Nile virus in southern Florida. J Med Entomol. 2005; 42: 134–141. https://doi.org/10.1093/jmedent/42.2.134 PMID: 15799522
31. Paull SH, Horton DE, Ashfaq M, Rastogi D, Kramer LD, Diffenbaugh NS, et al. Drought and immunity determine the intensity of West Nile virus epidemics and climate change impacts. Proc R Soc B. 2017; 284: 20162078. https://doi.org/10.1098/rspb.2016.2078 PMID: 28179512
32. Gehlke CE, Biehl K. Certain effects of grouping upon the size of the correlation coefficient in census tract material. J Am Stat Assoc. 1934; 29: 169–170.
33. Openshaw S, Taylor P. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. Statistical Application in the Spatial Sciences. London: Pion; 1979.
34. Fotheringham AS, Wong DW. The modifiable areal unit problem in multivariate statistical analysis. Environ Plan A. 1991; 23: 1025–1044.
35. Pearson RG, Dawson TP. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob Ecol Biogeogr. 2003; 12: 361–371.
36. Cohen JM, Civitello DJ, Brace AJ, Feichtinger EM, Ortega CN, Richardson JC, et al. Spatial scale modulates the strength of ecological processes driving disease distributions. Proc Natl Acad Sci. 2016; 113: E3359–E3364. https://doi.org/10.1073/pnas.1521657113 PMID: 27247398
37. Rueda L, Patel K, Axtell R, Stinner R. Temperature-dependent development and survival rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae). J Med Entomol. 1990; 27: 892–898. https://doi.org/10.1093/jmedent/27.5.892 PMID: 2231624
38. Allan BF, Langerhans RB, Ryberg WA, Landesman WJ, Griffin NW, Katz RS, et al. Ecological correlates of risk and incidence of West Nile virus in the United States. Oecologia. 2009; 158: 699–708. https://doi.org/10.1007/s00442-008-1169-9 PMID: 18941794
39. Andreadis TG, Anderson JF, Vossbrinck CR, Main AJ. Epidemiology of West Nile virus in Connecticut: a five-year analysis of mosquito data 1999–2003. Vector-Borne Zoonotic Dis. 2004; 4: 360–378. https://doi.org/10.1089/vbz.2004.4.360 PMID: 15682518
40. Bowden SE, Magori K, Drake JM. Regional differences in the association between land cover and West Nile virus disease incidence in humans in the United States. Am J Trop Med Hyg. 2011; 84: 234–238. https://doi.org/10.4269/ajtmh.2011.10-0134 PMID: 21292890
41. Brown H, Diuk-Wasser M, Andreadis T, Fish D. Remotely-sensed vegetation indices identify mosquito clusters of West Nile virus vectors in an urban landscape in the northeastern United States. Vector-Borne Zoonotic Dis. 2008; 8: 197–206. https://doi.org/10.1089/vbz.2007.0154 PMID: 18452400
42. Brown HE, Childs JE, Diuk-Wasser MA, Fish D. Ecologic factors associated with West Nile virus transmission, northeastern United States. Emerg Infect Dis. 2008; 14: 1539. https://doi.org/10.3201/eid1410.071396 PMID: 18826816
43. Brownstein JS, Rosen H, Purdy D, Miller JR, Merlino M, Mostashari F, et al. Spatial analysis of West Nile virus: rapid risk assessment of an introduced vector-borne zoonosis. Vector Borne Zoonotic Dis. 2002; 2: 157–164. https://doi.org/10.1089/15303660260613729 PMID: 12737545

44. DeFelice NB, Little E, Campbell SR, Shaman J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nat Commun. 2017; 8: 14592. https://doi.org/10.1038/ncomms14592 PMID: 28233783

45. DeFelice NB, Schneider ZD, Little E, Barker C, Caillouet KA, Campbell SR, et al. Use of temperature to improve West Nile virus forecasts. PLoS Comput Biol. 2018; 14: e1006047. https://doi.org/10.1371/journal.pcbi.1006047 PMID: 29522514

46. Diuk-Wasser MA, Brown HE, Andreadis TG, Fish D. Modeling the spatial distribution of mosquito vectors for West Nile virus in Connecticut, USA. Vector-Borne Zoonotic Dis. 2006; 6: 283–295. https://doi.org/10.1089/vbz.2006.6.283 PMID: 16989568

47. Gates MC, Boston RC. Irrigation linked to a greater incidence of human and veterinary West Nile virus cases in the United States from 2004 to 2006. Prev Vet Med. 2009; 89: 134–137. https://doi.org/10.1016/j.prevetmed.2008.12.004 PMID: 19185941

48. Myer MH, Campbell SR, Johnston JM. Spatiotemporal modeling of ecological and sociological predictors of West Nile virus in Suffolk County, NY, mosquitoes. Ecosphere. 2017; 8: e01854. https://doi.org/10.1002/ecs2.1854 PMID: 30147987

49. 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. https://doi.org/10.1016/j.scitotenv.2018.09.397 PMID: 30373059

50. Rochlin I, Harding K, Ginsberg H, Campbell S. Comparative analysis of distribution and abundance of West Nile and eastern equine encephalomyelitis virus vectors in Suffolk County, New York, using human population density and land use/cover data. J Med Entomol. 2008; 45: 563–571. https://doi.org/10.1603/0022-2585(2008)45[563:caodaa]2.0.co;2 PMID: 18533453

51. Rochlin I, Ginsberg HS, Campbell SR. Distribution and abundance of host-seeking *Culex* species at three proximate locations with different levels of West Nile virus activity. Am J Trop Med Hyg. 2009; 80: 661–668. PMID: 19346396

52. Rochlin I, Turbow D, Gomez F, Ninivaggi DV, Campbell SR. Predictive Mapping of Human Risk for West Nile Virus (WNV) Based on Environmental and Socioeconomic Factors. PLOS ONE. 2011; 6: e23280. https://doi.org/10.1371/journal.pone.0023280 PMID: 21853103

53. Tonjes DJ. Estimates of worst case baseline West Nile virus disease effects in a suburban New York county. J Vector Ecol. 2008; 33: 293–304. PMID: 19263849

54. Trawinski PR, Mackay DS. Identification of environmental covariates of West Nile virus vector mosquito population abundance. Vector-Borne Zoonotic Dis. 2010; 10: 515–526. https://doi.org/10.1089/vbz.2008.0063 PMID: 20482343

55. Young SG, Tullis JA, Cothren J. A remote sensing and GIS-assisted landscape epidemiology approach to West Nile virus. Appl Geogr. 2013; 45: 241–249.

56. Ciota AT, Kramer LD. Vector-virus interactions and transmission dynamics of West Nile virus. Viruses. 2013; 5: 3021–3047. https://doi.org/10.3390/v5123021 PMID: 24351794

57. Breiman L. Random forests. Mach Learn. 2001; 45: 5–32.

58. Python Software Foundation. Python Language Reference, version 2.7.12 (32-bit) [Internet]. 2016. Available: http://www.python.org

59. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available: https://www.R-project.org/

60. New York State Department of Health. New York Mosquito Trap Data. To access these data, contact: Bryon Backenson at bryon.backenson@health.ny.gov. 2018.

61. Connecticut Agricultural Experimental Station. Connecticut Mosquito Trap Data. These data contain sensitive information; for access contact Theodore Andreadis at theodore.andreadis@ct.gov. 2018.

62. NYSDOH. Communicable Disease Annual Reports and Related Information [Internet]. New York State Department of Health; 2018. Available: https://www.health.ny.gov/statistics/diseases/communicable/

63. Connecticut State Department of Public Health. West Nile Virus Statistics [Internet]. 2018. Available: https://portal.ct.gov/DPH/Infectious-Diseases/EEI/West-Nile-Virus-Statistics

64. Hayes EB, Sejvar JJ, Zaki SR, Lanciotti RS, Bode AV, Campbell GL. Virology, pathology, and clinical manifestations of West Nile virus disease. Emerg Infect Dis. 2005; 11: 1174. https://doi.org/10.3201/eid1108.050289b PMID: 16102303

65. Mostashari F, Bunning ML, Kitsutani PT, Singer DA, Nash D, Cooper MJ, et al. Epidemic West Nile encephalitis, New York, 1999: results of a household-based seroepidemiological survey. The lancet. 2001; 358: 261–264.

66. Busch MP, Wright DJ, Custer B, Tobler LH, Stramer SL, Kleinman SH, et al. West Nile virus infections projected from blood donor screening data, United States, 2003. Emerg Infect Dis. 2006; 12: 395. https://doi.org/10.3201/eid1203.051287 PMID: 16704775

67. Reiter P. A portable battery-powered trap for collecting gravid Culex mosquitoes. Mosq News. 1983; 43: 496–498.

68. Turell MJ, Dohm DJ, Sardelis MR, O'guinn 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. https://doi.org/10.1093/jmedent/42.1.57 PMID: 15691009

69. Turell MJ, O'Guinn M, Oliver J. Potential for New York mosquitoes to transmit West Nile virus. Am J Trop Med Hyg. 2000; 62: 413–414. PMID: 11037788

70. Simpson JE, Hurtado PJ, Medlock J, Molaei G, Andreadis TG, Galvani AP, et al. Vector host-feeding preferences drive transmission of multi-host pathogens: West Nile virus as a model system. Proc R Soc Lond B Biol Sci. 2012; 279: rspb20111282.

71. Kulasekera VL, Kramer L, Nasci RS, Mostashari F, Cherry B, Trock SC, et al. West Nile virus infection in mosquitoes, birds, horses, and humans, Staten Island, New York, 2000. Emerg Infect Dis. 2001; 7: 722–725. https://doi.org/10.3201/eid0704.010421 PMID: 11589172

72. Hamer GL, Kitron UD, Brawn JD, Loss SR, Ruiz MO, Goldberg TL, et al. *Culex pipiens* (Diptera: Culicidae): A Bridge Vector of West Nile Virus to Humans. J Med Entomol. 2008; 45: 125–128. https://doi.org/10.1603/0022-2585(2008)45[125:cpdcab]2.0.co;2 PMID: 18283952

73. Molaei G, Andreadis TG, Armstrong PM, Anderson JF, Vossbrinck CR. Host Feeding Patterns of Culex Mosquitoes and West Nile Virus Transmission, Northeastern United States. Emerg Infect Dis. 2006; 12: 468–474. https://doi.org/10.3201/eid1203.051004 PMID: 16704786

74. Spielman Andrew. Structure and Seasonality of Nearctic *Culex pipiens* Populations. Ann N Y Acad Sci. 2001; 951: 220–234. https://doi.org/10.1111/j.1749-6632.2001.tb02699.x PMID: 11797779

75. Bernard KA, Kramer LD. West Nile virus activity in the United States, 2001. Viral Immunol. 2001; 14: 319–338. https://doi.org/10.1089/08828240152716574 PMID: 11792062

76. Darsie R, Ward R. Identification and Geographical Distribution of the Mosquitoes of North America, North of Mexico. 1st ed. University Press of Florida, USA; 1981.

77. Stojanovich CJ. Illustrated key to common mosquitoes of northeastern North America. Atlanta, GA: Cullom and Gherter; 1961.

78. Means R. The Genus *Aedes* Meigen, with Identification Keys to Genera of Culicidae. N Y State Mus Bull. 1979; 430a: 1–221.

79. Means R. Mosquitoes of New York: Part II. Genera of Culicidae other than *Aedes*. N Y State Mus Bull. 1987; 430b: 1–180.

80. Carpenter SJ, La Casse WJ. Mosquitoes of North America (North of Mexico). Berkeley: Univ of California Press; 1955.

81. Chiang CL, Reeves WC. Statistical estimation of virus infection rates in mosquito vector populations. Am J Hyg. 1962; 75: 377–91. PMID: 13878878

82. Walter SD, Hildreth SW, Beaty BJ. Estimation of infection rates in populations of organisms using pools of variable size. Am J Epidemiol. 1980; 112: 124–128. https://doi.org/10.1093/oxfordjournals.aje.a112961 PMID: 7395846

83. Williams CJ, Moffitt CM. Estimation of pathogen prevalence in pooled samples using maximum likelihood methods and open-source software. J Aquat Anim Health. 2005; 17: 386–391.

84. Biggerstaff BJ. PooledInfRate Version 4.0: a Microsoft® Office\copyright Excel Add-In to compute prevalence estimates from pooled samples. Fort Collins, CO: Centers for Disease Control and Prevention; 2009.

85. Newman AJ, Clark MP, Craig J, Nijssen B, Wood A, Gutmann E, et al. Gridded ensemble precipitation and temperature estimates for the contiguous United States. J Hydrometeorol. 2015; 16: 2481–2500.

86. Menne MJ, Durre I, Korzeniewski B, McNeal S, Thomas K, Yin X, et al. Global historical climatology network-daily (GHCN-Daily), Version 3. NOAA Natl Clim Data Cent. 2012; https://doi.org/10.7289/V5D21VHZ

87. Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG. An overview of the global historical climatology network-daily database. J Atmospheric Ocean Technol. 2012; 29: 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

88. Clark MP, Slater AG. Probabilistic quantitative precipitation estimation in complex terrain. J Hydrometeorol. 2006; 7: 3–22.

89. Tebaldi C, Hayhoe K, Arblaster JM, Meehl GA. Going to the Extremes. Clim Change. 2006; 79: 185–211. https://doi.org/10.1007/s10584-006-9051-4

90. Park Williams A., Cook Benjamin I, Smerdon Jason E., Bishop Daniel A., Seager Richard, Mankin Justin S. The 2016 Southeastern U.S. Drought: An Extreme Departure From Centennial Wetting and Cooling. J Geophys Res Atmospheres. 2017; 122: 10,888–10,905. https://doi.org/10.1002/2017JD027523 PMID: 29780677

91. Ek M. B., Mitchell K. E., Lin Y., Rogers E., Grunmann P., Koren V., et al. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J Geophys Res Atmospheres. 2003;108. https://doi.org/10.1029/2002JD003296

92. Niu Guo-Yue, Yang Zong-Liang, Mitchell Kenneth E., Chen Fei, Ek Michael B., Barlage Michael, et al. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. J Geophys Res Atmospheres. 2011;116. https://doi.org/10.1029/2010JD015139

93. Xia Y, Mitchell K, Ek M, Cosgrove B, Sheffield J, Luo L, et al. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. J Geophys Res Atmospheres. 2012;117.

94. Lampman RL, Novak RJ. Oviposition preferences of *Culex pipiens* and *Culex restuans* for infusion-baited traps. J Am Mosq Control Assoc. 1996; 12: 23–32. PMID: 8723254

95. Jackson BT, Paulson SL, Youngman RR, Scheffel SL, Hawkins B. Oviposition preferences of *Culex restuans* and *Culex pipien*s (Diptera: Culicidae) for selected infusions in oviposition traps and gravid traps. J Am Mosq Control Assoc. 2005; 21: 360–365. https://doi.org/10.2987/8756-971X(2006)21[360:OPOCRA]2.0.CO;2 PMID: 16506560

96. Sauer JR, Niven DK, Hines JE, Ziolkowski DJ Jr, Fallon JE, Link WA. The North American Breeding Bird Survey, Results and Analyses 1966–2015. Version 2.07.2017 [Internet]. USGS Patuxent Wildlife Research Center, Laurel, MD; 2017. Available: https://www.mbr-pwrc.usgs.gov/bbs/bbs.html

97. Komar N, Panella NA, Langevin SA, Brault AC, Amador M, Edwards E, et al. Avian hosts for West Nile virus in St. Tammany Parish, Louisiana, 2002. Am J Trop Med Hyg. 2005; 73: 1031–1037. PMID: 16354808

98. Komar N, Dohm DJ, Turell MJ, Spielman A. Eastern equine encephalitis virus in birds: relative competence of European starlings (Sturnus vulgaris). Am J Trop Med Hyg. 1999; 60: 387–391. https://doi.org/10.4269/ajtmh.1999.60.387 PMID: 10466964

99. Farnsworth GL, Nichols JD, Sauer JR, Fancy SG, Pollock KH, Shriner SA, et al. Statistical approaches to the analysis of point count data: a little extra information can go a long way. Bird conservation implementation and integration in the Americas: Proceedings of the Third International Partners in Flight Conference. US Department of Agriculture, Forest Service General Technical Report PSW-GTR-191; 2005. pp. 736–743.

100. Manson S, Schroeder J, Van Riper D, Ruggles S. IPUMS National Historical Geographic Information System: Version 12.0 [Database]. [Internet]. Minneapolis: University of Minnesota; 2017. Available: https://doi.org/10.18128/D050.V12.0

101. DeGroote JP, Sugumaran R. National and regional associations between human West Nile virus incidence and demographic, landscape, and land use conditions in the coterminous United States. Vector-Borne Zoonotic Dis. 2012; 12: 657–665. https://doi.org/10.1089/vbz.2011.0786 PMID: 22607071

102. Homer C, Dewitz J, Yang L, Jin S, Danielson P, Xian G, et al. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogramm Eng Remote Sens. 2015; 81: 345–354.

103. EPA. EPA Facility Registry Service (FRS): Wastewater Treatment Plants. U.S. Environmental Protection Agency, Headquarters; 2013.

104. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2: 18–22.

105. Alexander D, Tropsha A, Winkler DA. Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model. 2015; 55: 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206 PMID: 26099013

106. Lachenbruch PA, Mickey MR. Estimation of Error Rates in Discriminant Analysis. Technometrics. 1968; 10: 1–11. https://doi.org/10.2307/1266219

107. Whittaker J. Model interpretation from the additive elements of the likelihood function. Appl Stat. 1984; 52–64.

108. Lawler JJ, Edwards TC Jr. A variance-decomposition approach to investigating multiscale habitat associations. The Condor. 2006; 108: 47–58.

109. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: Classification and Regression Training [Internet]. 2017. Available: https://CRAN.R-project.org/package=caret

110. Kvalseth TO. Cautionary Note about R2. Am Stat. 1985; 39: 279–285. https://doi.org/10.2307/2683704

111. Census Bureau U.S. 2010 Census County—Connecticut [Internet]. U.S. Department of Commerce, U.S. Census Bureau, Geography Division; 2010. Available: http://magic.lib.uconn.edu/magic_2/vector/37800/countyct_37800_0000_2010_s100_census_1_t.htm

112. Paz S, Albersheim I. Influence of warming tendency on *Culex pipiens* population abundance and on the probability of West Nile Fever outbreaks (Israeli case study: 2001–2005). EcoHealth. 2008; 5: 40–48. https://doi.org/10.1007/s10393-007-0150-0 PMID: 18648796

113. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010; 36: 1–13.

114. Nilsson R, Peña JM, Björkegren J, Tegnér J. Consistent feature selection for pattern recognition in polynomial time. J Mach Learn Res. 2007; 8: 589–612.

115. Ruiz MO, Tedesco C, McTighe TJ, Austin C, Kitron U. Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. Int J Health Geogr. 2004; 3: 8. https://doi.org/10.1186/1476-072X-3-8 PMID: 15099399

116. 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. https://doi.org/10.1186/1476-072X-6-10 PMID: 17352825

117. Harrigan RJ, Thomassen HA, Buermann W, Cummings RF, Kahn ME, Smith TB. Economic conditions predict prevalence of West Nile virus. PLoS One. 2010; 5: e15437. https://doi.org/10.1371/journal.pone.0015437 PMID: 21103053

118. Montgomery R. Age-related alterations in immune responses to West Nile virus infection. Clin Exp Immunol. 2017; 187: 26–34. https://doi.org/10.1111/cei.12863 PMID: 27612657

119. Dale MR, Fortin M-J. Spatial analysis: a guide for ecologists. Cambridge University Press; 2005.

120. Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography. 2007; 30: 609–628.

121. Pimentel RS, Niewiadomska-Bugaj M, Wang J-C. Association of zero-inflated continuous variables. Stat Probab Lett. 2015; 96: 61–67. https://doi.org/10.1016/j.spl.2014.09.002

122. Andreadis TG. The contribution of *Culex pipiens* complex mosquitoes to transmission and persistence of West Nile virus in North America. J Am Mosq Control Assoc. 2012; 28: 137–151.

123. Vitart F, Ardilouze C, Bonet A, Brookshaw A, Chen M, Codorean C, et al. The subseasonal to seasonal (S2S) prediction project database. Bull Am Meteorol Soc. 2017; 98: 163–173.