West Nile virus forecast model submission form
Email completed form to vbd-predict@cdc.gov

Team name: Barker's Lab at UC Davis

Team leader

Name	Institution	Email
Matteo Marcantonio	University of California Davis	matmarcantonio@ucdavis.edu

Other team members

Name	Institution	Email
Christopher M Barker	University of California Davis	cmbarker@ucdavis.edu
Pascale Stiles	University of California Davis	pcstiles@ucdavis.edu
Anna Kawiecki	University of California Davis	akawiecki@ucdavis.edu
Karen Holcomb	University of California Davis	kmholcomb@ucdavis.edu
Sarah Taysir Abusaa	University of California Davis	stabusaa@ucdavis.edu
Marisa Donnelly	University of California Davis	madonnelly@ucdavis.edu

Model description
Provide a brief summary of the model methods with sufficient detail for another modeler to understand the approach being applied. If multiple models are used, describe each model and how they were combined.

WNV neuroinvasive cases data were modeled through a spatio-temporal Bayesian-inference hurdle model with two likelihoods, a Binomial likelihood for the occurrence of WNV, and a truncated negative binomial for the number of WNV occurrences. The first part of the model presents a binary component that generates zeros and ones, where zero corresponds to the zero values and one corresponds to positive values. The second part of the model has a count component that generates non-zero values.

The number of occurrences in county *i* in year *j* can be zero or a positive number, and WNV occurrence *z* at location *i* and year *j* can be defined as:

$$z_{i,j} = \begin{cases} 1 & \text{if there is at least 1 WNV case} \\ 0 & \text{otherwise} \end{cases}$$

whereas the count of WNV occurrences $o_{i,j}$ can be defined as:

$$o_{i,j} = \begin{cases} NA & \text{if } z_{i,j} = 0 \\ o_{i,j} & \text{otherwise} \end{cases}$$

We used a Binomial model for the binary process $z_{i,j}$ and a zero-truncated Negative Binomial (NB) for the positive WNV count. The NB was chosen over the Poisson model family due to overdisperse WNV count. The two model likelihood are defined as:

$z_{i,j} \sim Binomial(p_{i,j}, n_{i,j})$

$o_{i,j} \sim NB(r_{i,j}, p_{i,j})$

In the binomial model, p_{ij} represents the probability of observing the occurrence of WNV cases given $n_{i,j}$ number of trials which we considered to be population older than 65 years. In the NB model, $r_{i,j}$ and $p_{i,j}$ represent the number of 0 cases observed (i.e., dispersion parameter) before observing o_{ij} number of cases given $p_{i,j}$ probability of observing 0 cases in each trial, respectively. An offset term equal to the natural logarithm of population over 5 years in county i (E_i) was set on the NB likelihood. Thus, we assumed that the detection of the virus is associated with the population at high risk (people >65), whereas the overall count of neuroinvasive WNV cases is associated with the total population.
The linear predictors for both models $logit(p_{i,j})$ and $log(p_{i,j})$ are defined considering a range of covariates within the fixed design matrix $X^T_{i,j}$. We assumed that the "random effect" component of the models are represented by a sum of terms considering the dependence of data respect to space i (county) and time j (year):

$$logit(p_{i,j}) \;||\; log(p_{i,j}/E_i) = X^T_{i,j}\beta + \alpha_1\gamma_j + \alpha_2 b_i$$

where γ_j is a temporal effect representing unknown features of j count that have temporal structure a priori. For γ_j we adopted a prior considering alike effects for two neighbor time points, which is a random walk of second order. The model term b represents Besag-York-Mollié (BYM) model which we applied for county (i) spatial grouping. The BYM model is a combination of an exchangeable area-specific random effect v, and the spatially structured area-specific effect, u (Besag model). Here, we used a scaled parameterization of the BYM model, BYM2, which has been shown to perform better than BYM. Both spatially and temporally structured effects are shared between the two likelihoods of the hurdle model, and the α's represent their respective scaling parameters.
The fixed effect matrix $X^T_{i,j}$ is composed by covariates which vary only in space (county), such as land use, bird species richness and relative abundance, or in space (county) and time (month), such as climatic conditions.
The model was implemented using integrated nested Laplace approximations (INLA) algorithm. We assumed zero mean Gaussian prior distributions $\beta \sim Normal(0,0.001)$ for all the regression parameters. We chose penalized complexity (PC) priors on hyperparameters τ_γ and $\tau_b \sim PC(U=1, \alpha=0.01)$, and $\Phi_{ui} \sim PC(U=0.5, \alpha=0.5)$, where τ's represent precision parameters and Φ_{ui} is the mixing parameter of the BYM2 model. PC is a prior distribution which, if data are not informative, shrinks the parameter estimate toward a "base model," that is, characterized by spatial overdispersion or by no spatial or temporal auto-regressive structure, preventing overfitting. We set informative priors $Normal(10, 1)$ on both the scaling parameters α as, the higher the probability of occurrence, the greater the number of occurrences.

Model selection was performed starting with an only random effect model and iteratively adding variables while checking whether the predictive ability of the model data improved. As criterion for model selection, we used the logscore of the Conditional Prediction Ordinate

(CPO). Variables were added following an importance order decided *a priori* and based on their "biological" proximity respect to WNV neuroinvasive cases: 1) temporally-matching climatic variables, 2) land use, 3) major bird hosts relative abundance and total bird richness, 4) 1-year lagged climatic variables.

Areas unsuitable for WNV transmission cycle were excluded from the spatial aggregation of climatic variables through a "raster mask". The raster mask was composed of areas which were classified as "arid hot and cold desert, polar tundra and polar frost" (Köppen-Geiger climate classification) and were not agricultural land use (NCLD 2016). Moreover, pixels whose July-August average temperature (for the period 2000-2018) was lower than 12°C were also excluded. All model covariates were centered (subtracting the mean) and scaled (dividing by the standard deviation).

Variables

List each variable used and its temporal relationship to the forecast. If multiple models are used, specify which enter into each model.

1. Average temperature of August (PRISM; NASA NEX RCP2.6 for prediction); matching time.
2. Average temperature of January (PRISM; NASA NEX RCP2.6 for prediction); matching time.
3. Average temperature of February (PRISM; NASA NEX RCP2.6 for prediction); matching time.
4. Cumulative precipitation of May (PRISM; NASA NEX RCP2.6 for prediction); matching time.
5. Cumulative precipitation of August (PRISM; NASA NEX RCP2.6 for prediction); matching time.
6. Percentage coverage of developed land use (NLCD); 2016.
7. Percentage coverage of forest land use (NLCD); 2016.
8. Relative abundance of *Carpodacus mexicanus*, house finch (North American Breeding Bird Survey); 2000-2018 average.
9. Total bird species richness (North American Breeding Bird Survey); 2000-2018 average.
10. Average temperature of April (PRISM) ; 1-year lag.
11. Cumulative precipitation of October (PRISM) ; 1-year lag.
12. Palmer drought severity index August (NOAA); 1-year lag.
13. Percentage uninsured population (US census); 2006-2018 average.
14. Population above 65 years (US Census Bureau); 2010. Number of trials in the Binomial likelihood.
15. Total population count (US Census Bureau); 2010. Offset term on the Negative-binomial likelihood.

Computational resources

Describe the programming languages and software tools that were used to write and execute the forecasts.

Data processing was performed in GRASS GIS v7.9 and QGIS v3.12.2 software.

The model was coded in the R language and run in R v6.3.3 through the R-INLA v20.3.17.
Publications Note whether the model was derived from previously published work and, if so, provide references.
Participation agreement By submitting these forecasts, the team agrees to abide by the project rules and data use agreements.

Team lead name	Date
Matteo Marcantonio	2020-04-30