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

Team name
University of Nebraska-Lincoln

Team leader

Name	Institution	Email
Andrew J. Tyre	UNL	atyre2@unl.edu

Other team members

Name	Institution	Email
Kelly H. Smith	UNL	ksmith2@unl.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.

We used the R package mgcv to fit generalized additive models with thin-plate splines for non-parametric modeling of distributed lags of drought and temperature data, using restricted maximum likelihood estimation with a log link and negative binomial distribution (Wood, 2011). Natural-log-transformed population was used as an offset variable to directly model cases per 100,000 people.

If there is something unique about a county or year that is not reflected in the covariates, then that county or year could have consistently higher or lower cases than expected. This intra-class correlation can occur whenever a sample unit is measured repeatedly, as we do with both counties (multiple years) and years (many counties) (Zuur, 2007). One approach to account for this correlation is to include random effects, coefficients specific to a unit that are assumed to come from a specific distribution (usually normal) with mean zero. Including random effects increases the computational complexity of a model, so as an alternative we estimated categorical fixed effects for year using sum-to-zero contrasts (also called effects coding). Using sum-to-zero contrasts we can interpret the remaining fixed effects as applying to an average year.

For each county and year, we created sets of lags of precipitation and temperature variables, working backward from July. Using July as the start of the lagged data, the July value was lag 0, June was lag 1, May, lag 2, and so on. For this prediction we used 24 months of precipitation and temperature data.

Our global model was

$$\ln(\lambda_{i,t}) = \beta_0 + f_1(temp_{i,t,m}) + \beta_1(CI_{i,t}) + \beta_{2t} + \ln\left(\frac{population_{i,t}}{100{,}000}\right) \quad (1)$$

$$y_{i,t} \sim NegBinom(\lambda_{i,t}, k)$$

where *i* = county of observation, *t* = year of observation, and *m* = months of lagged observations leading up to the start of the infection season, β_1 is the coefficient for cumulative incidence, and β_2 is a vector of sum-to-zero contrast coefficients to help account for unique temporal (year) characteristics. f_1 is a non-linear functional smoothing curve. β_0 is the intercept. λ is the expected rate of infection, and *k* is the overdispersion parameter for the negative binomial distribution.

Forecasting through 2020

We calculated CI through the provisional 2019 data, used 2019 population estimates, and filled in missing weather data (May through July of 2020) with county averages between 2000 and 2019 inclusive. We set the sum-to-zero factor for year to 2018, and then subtracted the estimated coefficient for 2018 from the fitted value prior to using the inverse link function. This produces a point forecast for an "average year". We obtained probability forecasts for each bin using the negative binomial distribution with the estimated value of k from the fitted model.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). *Analysing ecological data*. Springer.

Variables

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

1. CI – cumulative incidence per 100K - annual
2. County population in 100K – annual
3. Monthly average temperature by county – previous 24 months
4. Monthly total precipitation by county – previous 24 months

5.
6.
7.
8.
9.
10.

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

R Version 3.6.0
Required Packages:
flmtools 0.0.0.9000 https://github.com/atyre2/flmtools
mgcv 1.8-28.
Recommended Packages:
tidyverse
usmap

Publications
Note whether the model was derived from previously published work and, if so, provide references.

Smith, K.H., et al. (in review) Using climate to explain and predict West Nile Virus risk in Nebraska. Geohealth manuscript 2020GH000244.

Participation agreement
By submitting these forecasts, the team agrees to abide by the project rules and data use agreements.

Team lead name	Date
Andrew Tyre	5/24/2020