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

Team name
(Stanford) Cardinals

Team leader

Name	Institution	Email
Marissa Childs	Stanford University	marissac@stanford.edu

Other team members

Name	Institution	Email
Nicole Nova	Stanford University	nicolenova@stanford.edu
Mallory Harris	Stanford University	mharris9@stanford.edu
Devin Kirk	Stanford University	kirkd@stanford.edu
Morgan Kain	Stanford University	morganpkain@gmail.com
Erin Mordecai	Stanford University	emordeca@stanford.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.

Using Hessian based boosting (XGBoost) with the xgboost package in the R programming language, we fit boosted regression and classification trees for both count outcomes and categorical bin outcomes defined by the forecast challenge. For count outcomes we used poisson negative log-likelihood as an evaluation metric, and for categorical outcomes we used multiclass log loss.

1) For a grid of hyperparameters (learning rate = 0.2, 0.1, 0.05, 0.01; tree complexity = 4, 6, 8, 10; subsample ratio = 0.5, 0.65, 0.8; minimum child weight = 2, 5, 8), we used 5-fold cross-validation to identify the number of boosting rounds that minimized the evaluation metric on the test set. For this cross-validation, we used data from 2000–2016.
2) We then selected the 4 sets of hyperparameters with the best test cross-validation evaluation metric (2 each for count and categorical outcomes) and fit models to all of the 2000–2016 data, using the optimal number of boosting rounds identified in the cross-validation.
3) We predicted for 2017 and 2018 using the 4 fitted models, and evaluated out of sample model performance. Bin probabilities for the Poisson models were calculated as the sum of probabilities of the outcomes within a bin. Bin probabilities below exp(-10) were set to zero for both count and categorical models, with the remaining probability distributed to the bin otherwise identified as having the highest probability.
4) We scored the 2017–2018 predictions using the log score defined by the forecasting challenge. The boosted classification tree fit to categorical outcomes with a learning rate of 0.05, tree complexity 4, subsample ratio of 0.8, and minimum child weight of 2, and optimal number of boosting rounds of 350 performed best.
5) We fit a final model to all data 2000–2018 using these parameters, and predicted 2020 outcomes, setting bin probabilities to 0 as above when probability was otherwise less than

exp(-10). We also report point estimates based on the best fitting poisson model in the submission.

Variables

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

1. Proportion of the human population that lives in: rural, micropolitan, metropolitan. County level. Static 2010.
2. Mosquito distribution model estimates for *Culex tarsalis, quinquifaciatus, pipiens*. County level average and max. Static 2010.
3. One-year lagged annual number of horse cases of WNV. County level: 2006–2019. State level: 1999–2005.
4. County level average temperature-dependent relative R0 as a function of temperature for *Culex tarsalis, quinquifaciatus,* and *pipiens* over from May–August (one year lag and average of 1–3 year lag). Temperature derived from ERA5 daily aggregates climate reanalysis product.
5. County level average temperature from May–August (one year lag and average of 1–3 year lag). Temperature derived from ERA5 daily aggregates climate reanalysis product.
6. County level average temperature and number of days with mean temperature below 0 degrees C during December (one year lag). Temperature derived from ERA5 daily aggregates climate reanalysis product.
7. County level total precipitation during March–August (one year lag). Precipitation derived from ERA5 daily aggregates climate reanalysis product.
8. University of Idaho Palmer Drought Severity Index. County level averages for both January–March of year of prediction, and May–August of year of previous year.
9. Annual estimated county population (total size and density), one year lagged from the US Census Bureau.
10. Previously reported WNV neuroinvasive disease cases at the county level (one year lagged cases, total cases in the last 2 years, total cases in the last 2–5 years, number of years since WNV neuroinvasive disease cases were first reported in the county).
11. Previously reported WNV neuroinvasive disease cases at the state level (one year lagged cases, total cases in the last 2–5 years, number of years since WNV neuroinvasive disease cases were first reported in the state).
12. Previously reported WNV neuroinvasive disease cases in adjacent counties (total cases and average incidence) in the previous year.
13. Percent of county with urban, wetland, cropland, and forest land covers (two year lagged). Derived from MCD12Q1 MODIS Land Cover Type V6.

Computational resources

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

The models were fit with the xgboost package in the R programming language using the Stanford University computing cluster (Farmshare).

Publications

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

Model used relative R0 estimates as a function of temperature for three mosquito species from Mordecai et al (2019).

Mordecai, E.A., Caldwell, J.M., Grossman, M.K., Lippi, C.A., Johnson, L.R., Neira, M., Rohr, J.R., Ryan, S.J., Savage, V., Shocket, M.S., Sippy, R., Stewart Ibarra, A.M., Thomas, M.B. and Villena, O. (2019), Thermal biology of mosquito-borne disease. Ecol Lett, 22: 1690-1708. doi:10.1111/ele.13335

Participation agreement

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

Team lead name	Date
Marissa Childs	April 30, 2020