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

Team name
DyMSiM_NN

Team leader

Name	Institution	Email
Adrienne Kinney	The University of Arizona	Akinney1@math.arizona.edu

Other team members

Name	Institution	Email
Joceline Lega	The University of Arizona	lega@arizona.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 developed a system of two 1D convolutional neural networks to generate binned predictions. The first model is a binary classifier trained to predict whether a county will have WNV cases or not. An individual input sample for the binary classifier has shape 10x15. The 15 input variables are listed in the **Variables** section below. The 10 rows in a sample correspond to the previous ten years' worth of input variables for a given county, with each row corresponding to one year. Thus, predictions for 2020 for each county used the input variables for 2011-2020. The output for each sample is a 1x2 array with the first column listing the probability that there are 0 cases for the sample and the second column listing the probability that there is at least 1 case for the sample. The binary classifier's architecture is as follows: 3 1D convolutional layers, each with 64 filters, a kernel size of 3x1, and ReLu activation functions; 1 batch normalization layer with a ReLu activation function; 1 dropout layer with rate 0.25; 1 fully connected layer with a ReLu activation function; and 1 binary classification layer with a softmax activation function. The model is trained using an Adam optimizer and a weighted cross-entropy loss function where samples with at least 1 case are weighted x10 and samples with 0 cases are weighted x1. We train using batch sizes of 32 over 30 epochs. The training data consists of the available data from 2000-2015 for all counties in the United States, and we tested on available data from 2016-2018 for all counties in the United States.
The samples that had at least .75 probability of 0 are extracted from the system and given binned predictions as follows: 0.98 for 0-1 cases, 0.01 for 1-6 cases, 0.005 for 6-11 cases, 0.0025 for 11-21 cases, etc.

The samples that had less than .75 probability of 0 cases are fed into the second model. The second model is a categorical classifier trained to predict the probability of each of the bins (i.e. 0-1 cases, 1-6 cases, etc.). The input samples for this model are identical to the binary classifier. The categorical classifier's training process and architecture are also identical to the binary classifier with the exception of the last layer, which is a categorical classification layer rather than a binary classification layer. The weighted cross-entropy loss function weighted samples with 0 cases x0.1 and samples from all other bins x1. The training data are the samples with at least a .75 probability of 0 cases from the binary classifier. The output for each sample is a 1x15 array with the first column listing the probability of 0-1 cases, the second column listing the probability of 1-6 cases, etc.

Variables

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

1. USDA plant hardiness zone
2. US Census Bureau estimated population
3. Mean annual precipitation
4. Mean annual minimum temperature
5. Mean annual maximum temperature
6. Standard deviation of annual precipitation
7. Standard deviation of annual minimum temperature
8. Standard deviation of annual maximum temperature
9. Number of days where maximum temperature is below 13 degrees Celsius
10. Number of days where minimum temperature is below 13 degrees Celsius
11. Number of days where maximum temperature is between 13 and 21 degrees Celsius
12. Number of days where minimum temperature is between 13 and 21 degrees Celsius
13. Number of days where maximum temperature is between 21 and 38 degrees Celsius
14. Number of days where minimum temperature is between 21 and 38 degrees Celsius
15. Number of days where maximum temperature is above 38 degrees Celsius

Computational resources

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

The convolutional neural networks were written in Python using Keras with a Tensorflow backend.

Publications

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

The input variables were obtained from the following sources:

Variable 1:
USDA Plant Hardiness Zone Map, 2012. Agricultural Research Service, U.S. Department of Agriculture. Accessed from https://planthardiness.ars.usda.gov/.

Variable 2:
U.S. Census Bureau (2020). Population Estimates for Counties. Retrieved from https://www2.census.gov/programs-surveys/popest/datasets/

Variables 3-15:
Climate forcings in the MACAv1-METDATA were drawn from a statistical downscaling of global climate model (GCM) data from the Coupled Model Intercomparison Project 5 (CMIP5, Taylor et al. 2010) utilizing the Multivariate Adaptive Constructed Analogs (MACA, Abatzoglou and Brown, 2012) method with the METDATA(Abatzoglou, 2011) observational dataset as training data.

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

Team lead name	Date
Adrienne Kinney	7/31/2020