West Nile virus Forecasting Challenge

Since West Nile virus (WNV) was introduced in the United States in 1999, it has caused seasonal summer outbreaks that vary in size and location, with most areas having sporadic disease or intermittent outbreaks. No vaccine or specific treatment of WNV is currently available. Reducing mosquito exposure through vector control and personal protective behaviors are the primary forms of prevention. Predicting where and when WNV transmission will occur could help direct public health control efforts.

This is an ***open*** forecasting challenge to predict the annual number of West Nile neuroinvasive disease cases in the counties of the contiguous United Sates during the 2020 calendar year. One person from each participating team needs to register on https://predict.cdc.gov/. The target, data, forecasts, and evaluations are described below.

Timeline

- Project announcement and historical data release: March 2020.
- Initial forecast (and model description) due: April 30, 2020.
- Additional forecasts due (optional): May 31, June 30, and July 31, 2020.

Forecast target

The total number of West Nile virus (WNV) neuroinvasive disease cases (confirmed and probable following the WNV neuroinvasive disease case definition (https://ndc.services.cdc.gov/case-definitions/arboviral-diseases-neuroinvasive-and-non-neuroinvasive-2015/) reported to ArboNET (https://wwwn.cdc.gov/arbonet/Maps/ADB_Diseases_Map/index.html) from each county in the contiguous United States in 2022.

Data

WNV data. Data consist of total annual neuroinvasive disease counts for counties in the 48 states (including the District of Columbia) within the contiguous United States from 2000–2019. All data for 2000-2018 are “final”, meaning that the case counts reflect the reported totals for that year; data for 2019 are provisional and have not undergone complete data cleaning. The data will be provided in a standardized csv file after registration for the 2020 WNV Forecasting Challenge on https://predict.cdc.gov/.

Other data. Participants can use any other data source, like climate, weather, land use, mosquito surveillance, and human demographics, to develop their modeling framework.

Data use. All teams must agree to the following terms for use of the historic WNV data.

1. Access to ArboNET data is limited to team members named on the model submission form. The data provided will be treated as confidential and should not be provided to other persons. All other requests for access to ArboNET data should be directed to the CDC Arboviral Diseases Branch (dvbid2@cdc.gov). Comments or questions about the challenge should be directed to vbd-predict@cdc.gov.
2. The data are provided for the purpose of statistical reporting and analysis only, and may not be combined with other data or information for the purpose of matching records to identify individuals. Any information that could be used directly or indirectly to identify individuals will not be disclosed. If the identity of a person included in the data is discovered inadvertently, that information should not be disclosed or otherwise made public.
3. Analysis and reporting will be performed only on the variables and final data provided and should not be combined or compared to provisional data from the current or previous years.
4. Provisional data, other than that which is already publicly available, cannot be released.
5. The data provided, including any temporary or permanent files created from the ArboNET data, should be stored on a password protected computer. Copies of the data file(s) should not be made, even for back-up purposes. Hard copies of the data will be stored securely and shredded when they are no longer needed.

6. The team is responsible for obtaining Institutional Review Board (IRB) review of projects when appropriate.
7. ArboNET will be appropriately referenced in any publications or presentations that are derived from these data and a draft of the article or presentation will be provided to the CDC Arboviral Diseases Branch for review.

Forecasts

Teams need to submit a forecast for the total number of West Nile neuroinvasive disease cases in the counties in the contiguous United States for 2020. The forecast should include a point estimate and a probability distribution, using the supplied template (Appendix 1). Teams will also need to submit a Model Description (Appendix 2) that provides a brief description of their model and data used.

The initial forecasts and model descriptions will be due on April 30, 2020. Updated forecasts may be submitted by May 31, June 30, and July 31, 2020. Updated forecasts may use newly acquired data or updated methods, but are not required. Forecasts may be submitted and updated at any time prior to the due date.

Evaluation

Forecasts will be quantitatively evaluated using the logarithmic scoring rule (Appendix 3). An analysis will be conducted using the average logarithmic score to assess and compare forecasts across all counties at each time point. A joint manuscript will be prepared to disseminate findings on this comparison and the general performance of submitted forecasts. Participants may publish their own forecasts and results at any time.

Appendix 1 – Forecast format

Forecasts should be made in csv files matching the format in the template. Each csv should contain forecasts for all counties. For internal record keeping, teams may find it useful to include the forecast due date or submission date in the file name.

The forecast file includes a set of lines for each forecast representing binned probabilities for the range of outcomes. Each bin is defined by an inclusive minimum and a non-inclusive maximum, for example, the bin defined by `bin_start_incl` = 1 case and `bin_end_notincl` = 6 cases is assigned the probability that the number of cases is greater than or equal to 1 and less than 6 (i.e., 1, 2, 3, 4, or 5 cases are reported, 1 ≤ x < 6). The following set of bins are used for each forecast: 0 ≤ x < 1, 1 ≤ x < 6, 6 ≤ x < 11, …, 46 ≤ x < 51, 51 ≤ x < 101, 101 ≤ 151, 151 ≤ 201, 201 ≤ 1000. Each of these bins should have a probability between 0 and 1.0 (inclusive) and the sum of the probabilities assigned to each set of bins for one county should be 1.0. The forecast file also includes a line for each forecast representing the point prediction i.e., the most likely outcome for the specific target. A value for point prediction is required for submission; however, the point prediction will not be evaluated for this challenge.

Each row in the submission file represents a single bin and includes the following columns:

location: “State” and “County” as written in the data files with a hyphen: “State-County”. For example, “California-San Diego” or “Texas-Harris”. Do not include the word “County” and include spaces between words within the county or state name. The easiest way is to accomplish this is by matching the template available above to the input data.

target: “Total WNV neroinvasive disease cases”

type: “Bin” or “Point”. “Bin” specifies that the prediction is for a bin covering a range of possible outcomes. “Point” specifies the total predicted cases but will not be evaluated.

unit: “cases“

bin_start_incl: The inclusive lower bound for the bin, e.g., 0, 1, 6, 11, …, 151, 201. NA for point forecasts.

bin_end_notincl: The non-inclusive upper bound for the bin, e.g., 1, 6, 11, 16, …, 201, 1000. NA for point forecasts.

value: A probability for the number neuroinvasive disease cases in the bin defined by `bin_start_incl` and `bin_end_notincl`. This probability should be greater than or equal to 0 and less than or equal to 1.0 for all bins per county. Value for ‘point’ predictions can be zero or any positive integer and must be present but will not be evaluated for this challenge.

EXAMPLE: For forecast for a single location (Autauga County, Alabama) is shown below. The forecast consists of a point forecast (1 case) and a probability distribution that sums to 1. Note that a probability is assigned to each outcome, even if that probability is zero.

location	…	bin_start_incl	bin_end_notincl	value
Alabama-Autauga		NA	NA	1

Alabama-Autauga		0	1	0.133
Alabama-Autauga		1	6	0.276
Alabama-Autauga		6	11	0.067
Alabama-Autauga		11	16	0.067
Alabama-Autauga		16	21	0.067
Alabama-Autauga		21	26	0.067
Alabama-Autauga		26	31	0.067
Alabama-Autauga		31	36	0.067
Alabama-Autauga		36	41	0.067
Alabama-Autauga		41	46	0.067
Alabama-Autauga		46	51	0.067
Alabama-Autauga		51	101	0.0
Alabama-Autauga		101	151	0.0
Alabama-Autauga		151	201	0.0
Alabama-Autauga		201	1000	0.0

Appendix 2 - Model Description

Each team needs to submit a description of their model, using the following format, and submitted by email to the organizers (vbd-predict@cdc.gov). If updates are made to the model for subsequent forecasts, an updated model description should be provided to the organizers.

Team name

Team leader

Name	Institution	Email

Other team members

Name	Institution	Email

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.

Variables

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

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

Computational resources

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

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

Appendix 3 – Evaluation metric

Logarithmic Score

The logarithmic score is a proper scoring rule based on binned probability distributions. If *p* is the set of probabilities for a given forecast, and p_i is the probability assigned to the observed outcome *i*, the logarithmic score is:

$$S(p, i) = \ln (p_i)$$

For each forecast of each target, p_i will be set to the probability assigned to the single bin containing the observed outcome. Undefined natural logs (which occur when the probability assigned to the observed outcome was 0) will be assigned a value of -10.

References

Gneiting T and AE Raftery. (2007) Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 102(477):359-378. Available at: https://www.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Rosenfeld R, J Grefenstette, and D Burke. (2012) A Proposal for Standardized Evaluation of Epidemiological Models. Available at: http://delphi.midas.cs.cmu.edu/files/StandardizedEvaluation_Revised_12-11-09.pdf.