# West Nile virus forecasting challenge 2022

## Background

West Nile virus (WNV) is the leading cause of arboviral disease in the contiguous United States. An estimated 70–80% of WNV infections are asymptomatic; 20–30% of infected persons develop an acute systemic febrile illness and <1% of infected persons develop neuroinvasive disease (e.g., meningitis, encephalitis, or myelitis). Among patients with neuroinvasive disease, the case-fatality ratio is approximately 10%. Due to its severity and distinctive clinical features, diagnosis and reporting of neuroinvasive disease is considered more consistent and complete than that of non-neuroinvasive disease.

The first cases of WNV disease in the United States were identified in New York City in 1999; the virus subsequently spread westward, reaching the Pacific coast in 2003. Since then, WNV has caused seasonal outbreaks in summer and early fall, which vary in size and location. However, many areas in the continental United States either have no cases or only sporadic disease cases annually. 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.

### Challenge

This is an ***open*** forecasting challenge to predict the total number of WNV neuroinvasive disease cases in U.S. counties during the 2022 calendar year. Further information on the forecasting target, including counties, historical data, participation, and evaluation is available below.

### Timeline

- Project announcement and historical data release: February 2022.
- Initial forecast due: April 30, 2022.
- Additional forecasts due (optional): May 31, June 30, and July 31, 2022.

## Target

The total number of confirmed and probable West Nile virus (WNV) neuroinvasive disease cases (following the WNV neuroinvasive disease case definition) reported to ArboNET, the national arboviral surveillance system, from each county in the contiguous United States in 2022.

## WNV Surveillance Data

Data consist of total annual neuroinvasive disease counts for counties in the 48 states in the contiguous United States and the District of Columbia from 2000–2021*. The data will be provided in the following standardized csv file upon receipt of a signed Data Request form (see below). A detailed description of the fields included in this csv file is also included below.

"NeuroWNV_by_county_2000-2021_FULL.csv" contains all 3,108 counties in the contiguous U.S. and D.C. with zeroes for counties without reported neuroinvasive cases in a given year.

*2021 data initially will contain provisional data that is subject to change; final data are anticipated in mid-year 2022.

### Data use

All teams need to submit a completed Data Request form to vbd-predict@cdc.gov. The team lead should fill out item 1-3 on page 3 and date and sign page 4 to acknowledge the following limitations of ArboNET 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. Case-specific data cannot be released by county or any geographic unit smaller than a state.
5. Provisional data, other than that which is already publicly available, cannot be released.
6. 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.
7. The team is responsible for obtaining Institutional Review Board (IRB) review of projects when appropriate.
8. 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.

## Data dictionary

These data include annual counts of neuroinvasive West Nile virus disease cases reported to ArboNET by county. Each line represents all cases for a single county in a single year.

Description of Variables

**fips**: The five-digit FIPS code, which includes the two-digit state code and the three-digit county code. (Examples: 29099, 36005)

**county**: The name of the county where the case(s) resided, not including the word “county”. (Examples: Jefferson, Bronx)

**state**: The name of the state that reported the cases. (Examples: Missouri, New York)

**location**: State and county for reported cases in the format State-County, there are no spaces between the state and hyphen, and the hyphen and county; but there are spaces between words in a state or county name, can be used to match the template or to submission file. (Examples: Missouri-Jefferson, New York-Bronx)

**year**: The year (four-digits) the cases were reported. (Examples: 2002, 2014)

**count**: Number of West Nile virus neuroinvasive disease cases reported.

## Additional data notes

- In 2013, Bedford City, VA (FIPS 51515) was incorporated into Bedford County, VA (FIPS 51019). Historical data may have FIPS code 51515, while more recent data may have FIPS 51019. Only a prediction for Bedford County, VA (FIPS 51019) should be included in your forecast. Historical data for this county is reported as follows:
    - **fips**: 51019/51515
    - **county:** Bedford/Bedford City
    - **location**: Virginia-Bedford City/Bedford
- In 2015, Shannon County, SD (FIPS 46113) was renamed to Oglala Lakota County (FIPS 46102). Historical data may have FIPS code 46113 or the name Shannon County, while more recent data may use FIPS 46102 or the name Oglala Lakota County. Only a

prediction for Oglala Lakota County, SD (FIPS 46102) should be included in your forecast. Historical data for this county is reported as follows:

- **fips:** 46102/46113
- **county:** Oglala Lakota/Shannon
- **location:** South Dakota-Oglala Lakota/Shannon

## Participation

Work is still underway to setup an electronic submission system for 2022. In the interim, please email us at vbd-predict@cdc.gov and we will email all those interested when the participation process is finalized. In the interim, teams may proceed with the Data Use Agreement and preparing forecasts with the format described below.

Full participation requires:

1. Electronic submission of forecasts in the required format for all included counties by the initial deadline.
2. Submission of a model description document by email (see "*Model Methods*" section).
3. Submission of a signed data request document by email (see "*Data Use*").

### Forecast format

Forecasts should be made in csv files matching the format in the attached 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, $1 \leq x < 6$). The following set of bins are used for each forecast: $0 \leq x < 1$, $1 \leq x < 6$, $6 \leq x < 11$, …, $46 \leq x < 51$, $51 \leq x < 101$, $101 \leq x < 151$, $151 \leq x < 201$, $201 \leq x < 1000$. 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.

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

**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.

### Submission dates

The submission format will be automatically validated when the forecast is uploaded. Forecasts will be made for all of 2022 and initial forecasts will be due on April 30, 2022. Updated forecasts may be submitted by May 31, June 30, and July 31, 2022. 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

Final reported data for 2022 will be provided to all participants when available. 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.

### Eligibility

To be eligible, teams must:

1. Submit forecasts for every county included in the data ($n$ = 3,108).
2. Submit forecasts electronically prior to the deadline (April 30, 2022).
3. Submit a model description (see "*Participation*").

### Logarithmic Score

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/StandardizedEvaluationRevised12-11-09.pdf.


## Model Methods

The initial forecast submission should be accompanied by a completed Model Description form 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.