Title: Probabilistic Evaluation of Null Models for West Nile Virus in the United States

Running Head: Null Models for West Nile Virus

Authors: Alexander C. Keyel[1,2] and A. Marm Kilpatrick[3]

[1] Division of Infectious Diseases, Wadsworth Center, New York State Department of Health, Albany, NY, United States of America

[2] Department of Atmospheric and Environmental Sciences, University at Albany, SUNY, Albany, NY, United States of America

Contact information: akeyel@albany.edu; +1 518-442-4556

[3] Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA

Abstract

Null models provide a useful baseline for the development of new models. A variety of options for null models exist. These options have become more sophisticated with the advent of probabilistic modeling approaches. Here, we evaluate 10 different null models for West Nile virus, a primarily mosquito-borne disease introduced to the United States in 1999. The Historical Null was significantly better than all models other than the Negative Binomial. We recommend the use of either of these models as a baseline when developing new models to predict spatial and temporal dynamics of West Nile virus at the county-annual scale. We expect these results to be scale-dependent, and a future direction is to examine performance of null models at finer spatial and temporal scales.

Introduction

A classic definition of Ecology is the study of the distribution and abundance of organisms (Andrewartha & Birch 1954). Large data sets and new modeling and statistical techniques have made it possible to predict distributions and abundance into the future in the new discipline of Ecological Forecasting

(Dietze 2017). An important and early subdiscipline of Ecological Forecasting is predicting disease outbreaks (e.g., Ong *et al.* 2010; LaDeau *et al.* 2011). The use of Null models to provide a baseline in the absence of specific mechanisms has a long history in ecology (Gotelli & Graves 1996). Here, a null model is "a pattern-generating model that is based on randomization of ecological data or random sampling from a known or imagined distribution… designed to produce a pattern that would be expected in the absence of a particular ecological mechanism" (Gotelli & Graves 1996). Such baselines are important, as in some cases, predictive models may appear good (i.e. high accuracy), but may be no better than a null model (Olden *et al.* 2002). For example, when dealing with rare events, a null model that predicts the event to never occur may have high accuracy, but does not provide any information about the process being studied (Olden *et al.* 2002).

There has been a recent shift in ecological forecasting towards the use of probabilistic models (Dietze 2017). In a probabilistic model, rather than evaluating a single point forecast, a model's forecast is evaluated on the basis of a distribution of predicted outcomes (Fig. 1). The Continuous Ranked Probability Score is a useful method for scoring probabilistic forecasts (see Methods for a detailed justification), as it can score a distribution of predicted values relative to an observed value, and can accommodate point and probabilistic forecasts. With the development of probabilistic forecasts, it is important to consider a suite of probabilistic null models for model evaluation.

West Nile Virus (WNV) provides an excellent system in which to examine null models in a probabilistic context. WNV is a flavivirus that cycles between mosquito and avian populations (Work *et al.* 1955; Komar *et al.* 2003). WNV was introduced to the United States (US) in 1999 (Lanciotti *et al.* 1999) and rapidly spread to the conterminous US (CDC 2019a). As a nationally notifiable disease, long-term data sets (up to 20 years) exist on human cases (CDC 2019b). WNV has been extensively studied from a

predictive standpoint (LaDeau *et al.* 2011; Manore *et al.* 2014; Barker 2019). Prior studies of WNV have varied in their choice of null model, with many not including a null model for comparison.

We tested 10 null models for WNV in a probabilistic framework, including a novel historical null model based on previous cases. The study aims to provide guidance on null model selection and performance in disease forecasting for cumulative human cases with county-level resolution.

Methods

We considered 10 null models in our comparison. Data analyses were performed primarily in R (R Core Team 2017), with ranks assigned manually in MS Excel. Models were evaluated with the CRPS. Two common methods for evaluating probabilistic forecasts with continuous data are the Continuous Ranked Probability Score and the Logarithmic Score (Matheson & Winkler 1976; Hersbach 2000; Wilks 2011). Both methods are proper scoring rules. A proper scoring rule means the metric gives a better score to accurate forecasts over the long run, and cannot be 'gamed' to improve the forecast (see Bracher 2019 for an example where a scoring rule could be "gamed"). The Continuous Ranked Probability Score (CRPS) can be approximated based on samples from a distribution (Jordan *et al.* 2019; Bracher *et al.* 2021). The CRPS differs from the Logarithmic Score in that forecasts are scored based on the distance from each predicted probability to the observation. Thus, scores are better for models that are closer to the observed value. In contrast, the Logarithmic Score is based on bins, and an observation is either within a bin or outside of a bin, with no consideration of how far outside the bin the prediction was, or how close to the observed value within the bin the prediction was. Here, the CRPS was used, as proximity to the observation was considered an important property, and it avoided an arbitrary selection of bins. Code for running the null models is available via the dfmip package on GitHub (www.github.com/akeyel/dfmip/R/NullModels.R).

Always Absent Null (non-probabilistic): WNV is always predicted to be absent (i.e., the model always predicts 0). This model serves for a baseline for putting model errors into context. Any model that does not outperform this null model should not be considered as providing any useful information, regardless of accuracy statistics.

Mean Value Null (non-probabilistic): WNV is predicted to have its mean value of cases. As the CRPS scores based on distance from the value, fractional cases are allowed to be predicted under this model.

Stratified Mean Value Null (non-probabilistic): As in the mean value null, but stratified by county.

Prior Year Null (non-probabilistic): The results from the prior year are used to predict the current year (Smith *et al.* 2020). This null has a weakness in that avian seroprevalence often increases following a widespread WNV outbreak, and high avian immunity decreases WNV activity (Kwan *et al.* 2012). This model also is unable to take advantage of the rich historical data for a location.

Historical Null: Prior observations for the location are sampled 100 times with replacement, creating a probabilistic distribution of future cases. The null model is expected to perform better with increasing data availability, and expected to perform worse with increasing temporal and spatial resolution. I.e., as cases become rare, the model approaches the Always Absent Null.

Random Incidence Null: The disease incidence is calculated for the entire region of interest (i.e. the United States). A random draw was taken for every person in each county, if the random number was less than the incidence, then that was considered a case, and cases were summed for each county. This

process was repeated 100 times for each county to generate a probabilistic distribution. This null is not expected to perform particularly well, as it is known that there are regional hotspots for WNV (CDC 2019a).

Stratified Random Incidence Null: As in the random incidence null model, but incidence is calculated stratified by county, allowing the model to correct for local hotspots of WNV.

Negative Binomial Null: Human cases from each county were used to estimate a negative binomial distribution (mean and dispersion parameter) using the MASS package in R (Venables & Ripley 2002). Random draws (n = 100) for each county were then taken based on the estimated mean and dispersion parameter.

Autoregressive AR1: An autoregressive model was fit using the arima function in R (Ripley 2002). The model produced a mean estimate for the next time step and a normally-distributed error around that mean. We sampled the normal distribution around the mean 100 times. All values less than zero were assigned to 0, as negative human cases are not possible.

Uniform Null: Random draws (n = 100) were taken from a uniform distribution bounded by 0 to the maximum number of cases observed in a county.

Data set

We evaluated the null models using the CDC Neuroinvasive WNV Cases (Source: ArboNET, Arboviral Diseases Branch, Centers for Disease Control and Prevention, Fort Collins, Colorado, contact the CDC for data access). This is a national data set on WNV cases from the United States from 2000 - 2019. We

limited the data to neuroinvasive cases, as these are expected to be more standard across different states, due to the severity of the disease. Data were restricted to states that had at least one neuroinvasive case of West Nile virus, to avoid artificially zero-inflating the training data, leading to an increase in sample size over time (S1 Table, S2 Figure). Population data for the incidence-based null models came from the United States Census bureau (US Census Bureau 2017).

Results

The Negative Binomial model was the best model in 8 of the 16 years considered (mean rank 1.8, mean CRPS 0.298), with the Historical Null best in 5 years (mean rank 1.9, mean CRPS 0.291), the Always Absent model best in 2 years (mean rank 4.75, mean CRPS 0.385), and the stratified incidence model best in one year (mean rank 3.4, mean CRPS 0.324). Beginning in 2012, the Negative Binomial was the best model in 7 of the 8 years. There was no significant difference between the top two null models (historical and negative binomial nulls, Fig. 2). After correction for multiple comparisons (Holm 1979), the Negative Binomial model was not significantly better than the Always Absent model (Paired Wilcoxon test, unadjusted $p = 0.008$, Holm-adjusted significance threshold = 0.004, Fig. 2). Five null models performed worse on average than always predicting West Nile virus to be absent (Fig. 2). Computationally, the Always Absent and the Mean Value models were the fastest (<1 ms), while the Stratified and Random Incidence models were the slower than all the other models by more than an order of magnitude (> 16 minutes, S3 Table).

Case studies are presented for three counties: Cook, IL, Harris, TX, and Barbour, AL (Table 1, Fig. 3). Unlike in the comparison above, training was restricted to 2002 – 2012 and testing to 2013 - 2019 (the analysis above used all preceding data to predict the next year). The negative binomial fit for Cook County is poor (Fig. 1a), leading to an improved CRPS score for the Historical Null (Table 1). Both models

performed similarly for Harris and Barbour Counties. The Always Absent model scored slightly better than the Historical Null and the Negative Binomial in Barbour due to an absence of reported cases during the test data set (Table 1).

Discussion

We found the historical null and negative binomial models to be the best null models based on the CRPS scoring metric. The negative binomial model is able to produce smoother predictions than the historical null, by using the assumed distribution to fill data gaps. It is possible that the early poorer performance is due to insufficient training data, as this model was the best model 7 out of 8 times in the later part of the data series. The Negative Binomial distribution has a long history of use (e.g., Quenouille 1949), and the Negative Binomial Model had the best mean rank, and second lowest CRPS score (see S4 Table for the best model for each year). It was surprising that the Negative Binomial model was not significantly better than the Always Absent model, when the Historical Null was, given the better average ranked performance of the Negative Binomial. However, the Wilcoxon signed-rank test tests for deviations in a symmetric distribution of values around zero. There are 4 instances when the Always Absent model scored better than the Negative Binomial, compared to 2 for the Historical Null, leading to a greater deviation from the null distribution for the Historical Null model (S5 Figure). A future analysis with a planned, a priori contrast would likely reveal a significant difference between this pair of models. Importantly, the Historical Null finished ahead of the Negative Binomial in the two years when the Always Absent model ranked first, and in 2008, the Always Absent model finished ahead of the Negative Binomial, and these poor rankings likely made the difference between a near-significant difference and a significant difference. The Historical Null model does not smooth out the distribution, relying instead on historical values. However, a smoother distribution is less critical with the CRPS scoring method, where errors are not a binary right/wrong (e.g., as in the logarithmic scoring method), but are scored

169 based on their distance from the observed value. As a consequence, the Historical Null model scored the
170 best on the basis of lowest CRPS score. The uniform null model performed quite poorly in this study. The
171 Prior Year null (used in Smith *et al.* 2020) and the Mean Value (used in Keyel *et al.* 2019 in a non-
172 probabilistic context) to be poor choices of null models for WNV. The AR1 model also performed poorly.
173 It is likely that the autocorrelation structure is not properly captured with a single lag. While measles
174 shows a strong biennial cycle (Grenfell *et al.* 2001), transmission of WNV is much more complicated,
175 cycling between mosquitoes and primarily avian hosts (Work *et al.* 1955). Also, using a normal error
176 structure frequently produced negative values, suggesting that a different error structure (e.g., negative
177 binomial) might be more appropriate. This model warrants further attention as a potential null model
178 under different parameterizations. The strong performance of the Always Absent null is a reminder that
179 basic accuracy statistics for rare events can appear high when true absences are scored equally with true
180 positives. Indeed, during low WNV years (2004, 2009), this model was sometimes the top model (S4
181 Table). This is also a caution – just because a model performs well in any single year does not mean that
182 it is the best model for prediction purposes. Evaluation must take place over multiple years in order to
183 separate good performance by predicting one possible state outcome with high confidence from models
184 that captures the actual probability distribution. We believe that Null Models will be useful for cross-
185 regional comparisons of WNV (e.g., Keyel et al. *unpublished manuscript*). The county-annual scale
186 examined here is of academic interest, but is not the most useful scale for vector control and public
187 health responses (Keyel et al. *unpublished manuscript*). This research could be expanded by a study
188 evaluating null models across multiple spatial and temporal scales, especially focusing on the fine
189 temporal and spatial scales of interest to vector control and public health agencies. For example, a
190 recent forecasting challenge for dengue virus used a Uniform null and an Autoregressive Null for within-
191 season forecasts (Johansson *et al.* 2019). It would also be of interest to examine the degree to which
192 different counties contribute to the CRPS score. Specifically, it would be of interest to determine to what

degree the CRPS scores are driven by a few high-case counties compared to the contributions of the large number of counties that only infrequently have WNV cases. This could also be examined on a regional basis.

Conclusions

Including null models in ecological forecasting challenges are critically important. Probabilistic forecasts are evaluated differently than point forecasts, and therefore the options for null models are more varied. For infectious disease models, we recommend using a Negative Binomial model or a Historical Null model, as these were the top two models. We recommend the Always Absent model if events are rare (<50% of records), as this provides a baseline for accuracy. Conversely, for systems where events are common, an always present or mean value null are expected to perform better.

Acknowledgements

We thank Luis F. Chaves for constructive discussion. This publication was supported by cooperative agreement 1U01CK000509-01, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

References

Andrewartha, H.G. & Birch, L.C. (1954). *The distribution and abundance of animals.* University of Chicago press.

Barker, C.M. (2019). Models and Surveillance Systems to Detect and Predict West Nile Virus Outbreaks. *J Med Entomol*, 56, 1508–1515.

Bracher, J. (2019). On the multibin logarithmic score used in the FluSight competitions. *Proc Natl Acad Sci USA*, 116, 20809.

Bracher, J., Ray, E.L., Gneiting, T. & Reich, N.G. (2021). Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17, e1008618.

CDC. (2019a). Final Cumulative Maps & Data for 1999-2018.

CDC. (2019b). Nationally notifiable arboviral diseases reported to ArboNET: Data release guidelines.

Dietze, M. (2017). *Ecological Forecasting*. Princeton University Press.

Gotelli, N.J. & Graves, G.R. (1996). *Null models in ecology*.

Grenfell, B.T., Bjørnstad, O.N. & Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414, 716–723.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Johansson, M.A., Apfeldorf, K.M., Dobson, S., Devita, J., Buczak, A.L., Baugher, B., *et al.* (2019). An open challenge to advance probabilistic forecasting for dengue epidemics. *Proceedings of the National Academy of Sciences*, 116, 24268–24274.

Jordan, A., Krüger, F. & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, 90, 1–37.

Keyel, A.C., Elison Timm, O., Backenson, P.B., Prussing, C., Quinones, S., McDonough, K.A., *et al.* (2019). Seasonal temperatures and hydrological conditions improve the prediction of West Nile virus infection rates in Culex mosquitoes and human case counts in New York and Connecticut. *PLOS ONE*, 14, e0217854.

Komar, N., Langevin, S., Hinten, S., Nemeth, N., Edwards, E., Hettler, D., *et al.* (2003). Experimental infection of North American birds with the New York 1999 strain of West Nile virus. *Emerging infectious diseases*, 9, 311.

Kwan, J.L., Kluh, S. & Reisen, W.K. (2012). Antecedent Avian Immunity Limits Tangential Transmission of West Nile Virus to Humans. *PLOS ONE*, 7, e34127.

LaDeau, S.L., Glass, G.E., Hobbs, N.T., Latimer, A. & Ostfeld, R.S. (2011). Data–model fusion to better understand emerging pathogens and improve infectious disease forecasting. *Ecological Applications*, 21, 1443–1460.

Lanciotti, R.S., Roehrig, J.T., Deubel, V., Smith, J., Parker, M., Steele, K., *et al.* (1999). Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science*, 286, 2333–2337.

Manore, C.A., Davis, J.K., Christofferson, R.C., Wesson, D.M., Hyman, J.M. & Mores, C.N. (2014). Towards an early warning system for forecasting human West Nile virus incidence. *PLoS Currents*, 6.

Matheson, J.E. & Winkler, R.L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002). Predictive Models of Fish Species Distributions: A Note on Proper Validation and Chance Predictions. *null*, 131, 329–336.

Ong, J.B.S., Mark, I., Chen, C., Cook, A.R., Lee, H.C., Lee, V.J., *et al.* (2010). Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PloS one*, 5.

Quenouille, M.H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 5, 162–164.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B.D. (2002). Time series in R 1.5.0. *R News*, 2, 2–7.

Smith, K.H., Tyre, A.J., Hamik, J., Hayes, M.J., Zhou, Y. & Dai, L. (2020). Using Climate to Explain and Predict West Nile Virus Risk in Nebraska. *GeoHealth*, 4, e2020GH000244.

US Census Bureau. (2017). Intercensal estimates of the resident population for counties and states: April 1, 2000 to July 1, 2010. Suitland, MD: US Census Bureau. Retreived from: https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-counties.html.

Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*. Fourth. Springer, New York.

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Work, T.H., Hurlbut, H.S. & Taylor, R. (1955). Indigenous Wild Birds of the Nile Delta as Potential West Nile Virus Circulating Reservoirs. *The American Journal of Tropical Medicine and Hygiene*, 4, 872–888.

Table 1. A breakdown of the contribution of each year to the CRPS score for the Always Absent, Historical Null, and Negative Binomial models for three example counties. The CRPS value for the Always Absent model is also the number of WNV cases for that year. The CRPS becomes the mean absolute error when applied to a point estimate and the Always Absent model always predicts 0 cases.

	Cook, IL			Harris, TX			Barbour, AL		
Year	**AA**	**HN**	**NB**	**AA**	**HN**	**NB**	**AA**	**HN**[1]	**NB**[1]
2013	46	15.9	16.1	6	13.4	14.7	0	0.008	0.008
2014	23	8.3	14.6	103	57.9	55.2	0	0.008	0.008
2015	21	8.2	14.8	28	5.4	5.5	0	0.008	0.008
2016	65	24.9	21.9	18	5.3	7.0	0	0.008	0.008
2017	41	14.1	15.3	14	7.1	9.0	0	0.008	0.008
2018	76	31.5	26.7	31	6.3	5.8	0	0.008	0.008
2019	9	11.3	19.4	2	16.7	18.3	0	0.008	0.008
Mean	40.1	16.3	18.4	28.9	16.0	16.5	0	0.008	0.008

[1] By chance, these two models predicted the same distribution of zero and one case occurrences. Different random draws would cause these scores to vary slightly.

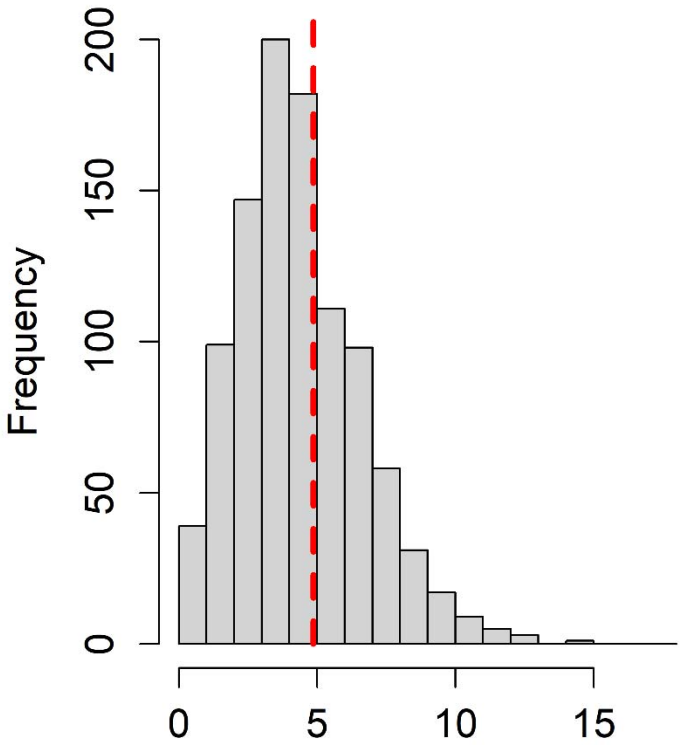


Fig. 1. A probabilistic forecast can convey substantially more information than a point forecast. Here, a dashed line indicates the point estimate, while the histogram of 1000 samples of a negative binomial distribution with $\mu = 5$ (mean) and size = 100 (dispersion parameter) show the expected range of possible outcomes.

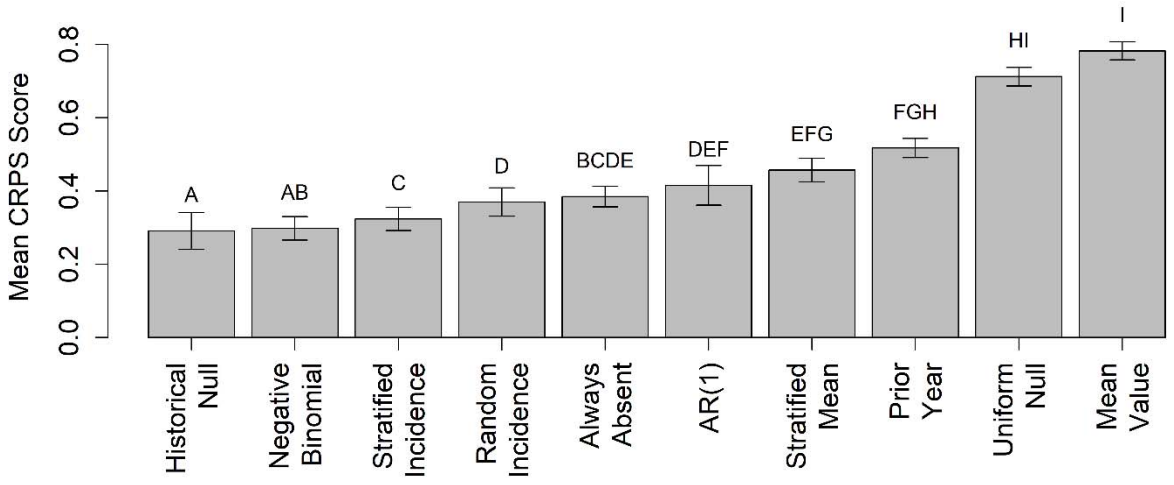


Fig. 2. Continuous Ranked Probability Scores for 2004 – 2019 for 10 null models. The Historical Null and Negative Binomial models were among the best null models, and were not significantly different from one another based on paired non-parametric Wilcoxon tests in R (R Core Team 2017) with a Holm correction for multiple pairwise comparisons (Holm 1979). Sample sizes changed between years from (553 counties in 2004 to 3108 beginning in 2015, see S1 Table and S2 Figure), as counties in states without a recorded case of WNV were excluded for those years. Note that error bars show standard error in CRPS scores, while the statistical test was performed on the basis of paired model differences (see S5 Figure for the distributions of differences for Historical Null, Negative Binomial and Always Absent model pairs).

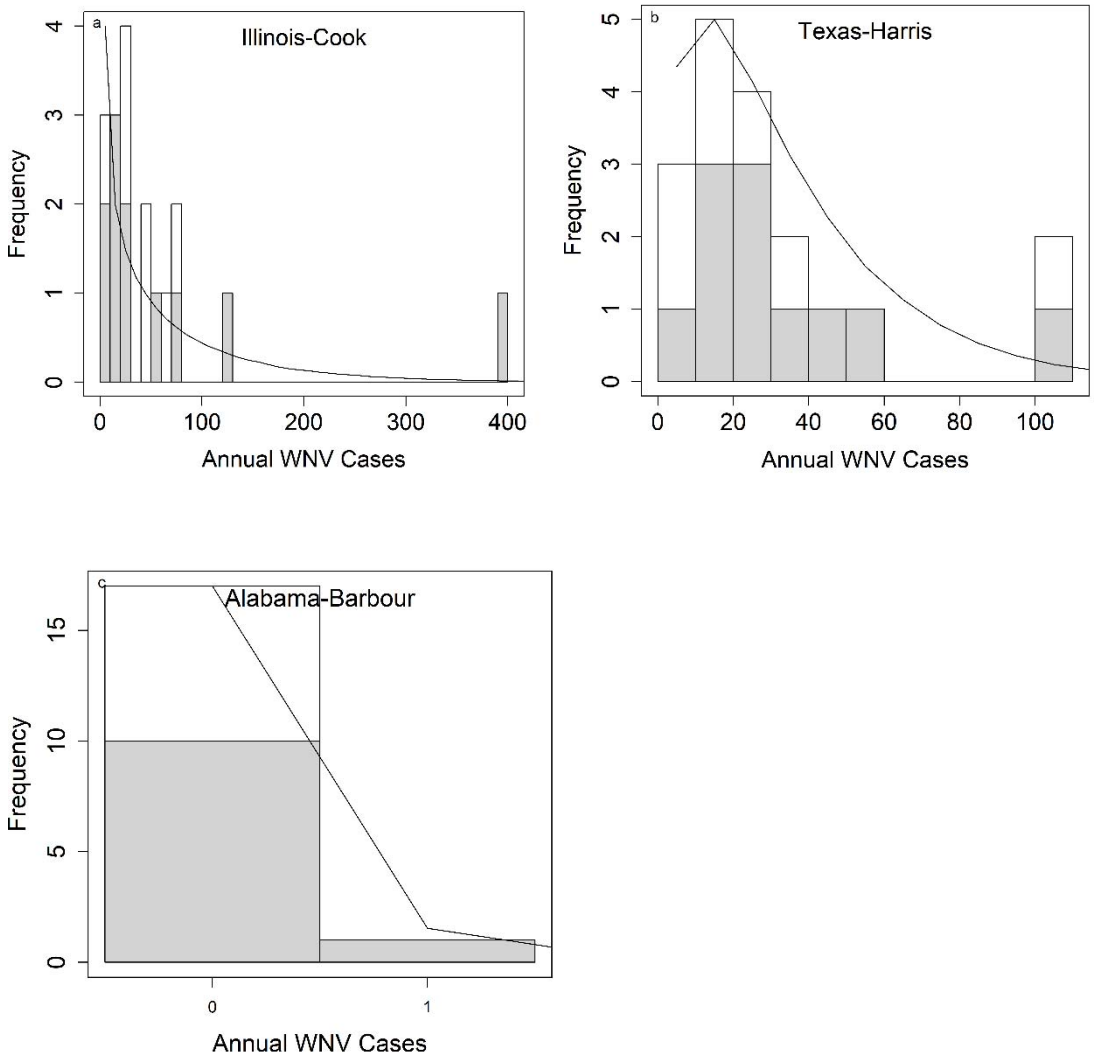
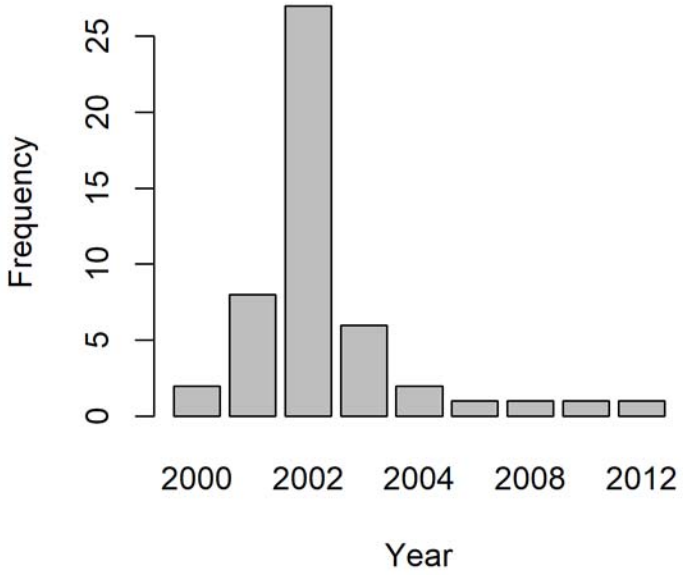


Fig. 3. Training (shaded, 2002 - 2012) and testing data (unshaded, 2013 - 2019) with a negative binomial fit based on the training data (line) for three selected counties: a) Cook County, IL, b) Harris County, TX, and c) Barbour County, AL. Cook and Harris were selected as they are among the counties with the most human cases of WNV. Barbour was selected to show a county with few WNV cases.

S1 Table. Number of counties used for the null model analysis by year.

Year	Sample Size
2004	553
2005	2829
2006	2930
2007	3003
2008	3039
2009	3039
2010	3039
2011	3078
2012	3078
2013	3078
2014	3092
2015	3108
2016	3108
2017	3108
2018	3108
2019	3108



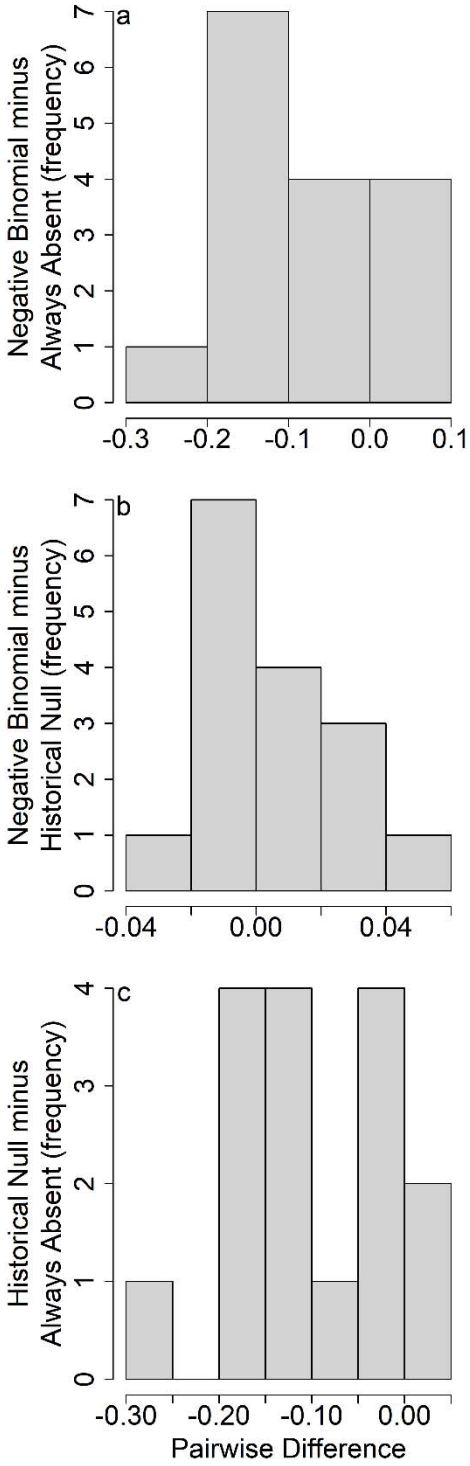
S2 Figure. The frequency of different years of first neuroinvasive WNV cases. Data from each state were excluded prior to at least one neuroinvasive case in the state to prevent zeros from WNV had not yet arrived from influencing the modeled fits. States on the tails of the distribution are: New York and New Jersey in 2000, Idaho and Utah in 2004, Oregon in 2005, Washington in 2008, Vermont in 2011, and Maine in 2012. Note that these are strictly based on neuroinvasive human cases in the CDC ArboNET database starting in 2000, so prior WNV activity in each state is possible.

S3 Table. Average time to run each null model for one year (restricted to 2015– 2019 to maintain a constant sample size).

Model	Elapsed Time (s)
Always Absent	<<1
Mean Value	<<1
Prior Year	10
Stratified Mean	12
Negative Binomial	14
Historical	14
Uniform	24
Autoregressive AR1	28
Stratified Incidence	990
Random Incidence	1118

S4 Table. Best model for each year based on CRPS

Year	Best Model	CRPS score (Gap to next model)
2004	Always Absent	0.316 (0.015)
2005	Stratified Incidence	0.331 (0.015)
2006	Negative Binomial	0.325 (0.002)
2007	Historical Null	0.306 (0.005)
2008	Historical Null	0.219 (0.007)
2009	Always Absent	0.118 (0.038)
2010	Historical Null	0.179 (0.024)
2011	Historical Null	0.146 (0.012)
2012	Negative Binomial	0.635 (0.000)
2013	Negative Binomial	0.268 (0.004)
2014	Negative Binomial	0.316 (0.002)
2015	Negative Binomial	0.308 (0.005)
2016	Negative Binomial	0.259 (0.001)
2017	Negative Binomial	0.285 (0.005)
2018	Negative Binomial	0.344 (0.006)
2019	Historical Null	0.173 (0.013)



S5 Figure. CRPS differences for a) Negative Binomial and Always Absent model, b) Negative Binomial and Historical Null model, and c), Historical Null and Always Absent models.