Artificial Intelligence Predicts and Explains West Nile Virus Risks Across Europe: Extraordinary Outbreaks Determined by Climate and Local Factors

Supplement:

cross-model comparability and the impact of scale-dependent variability

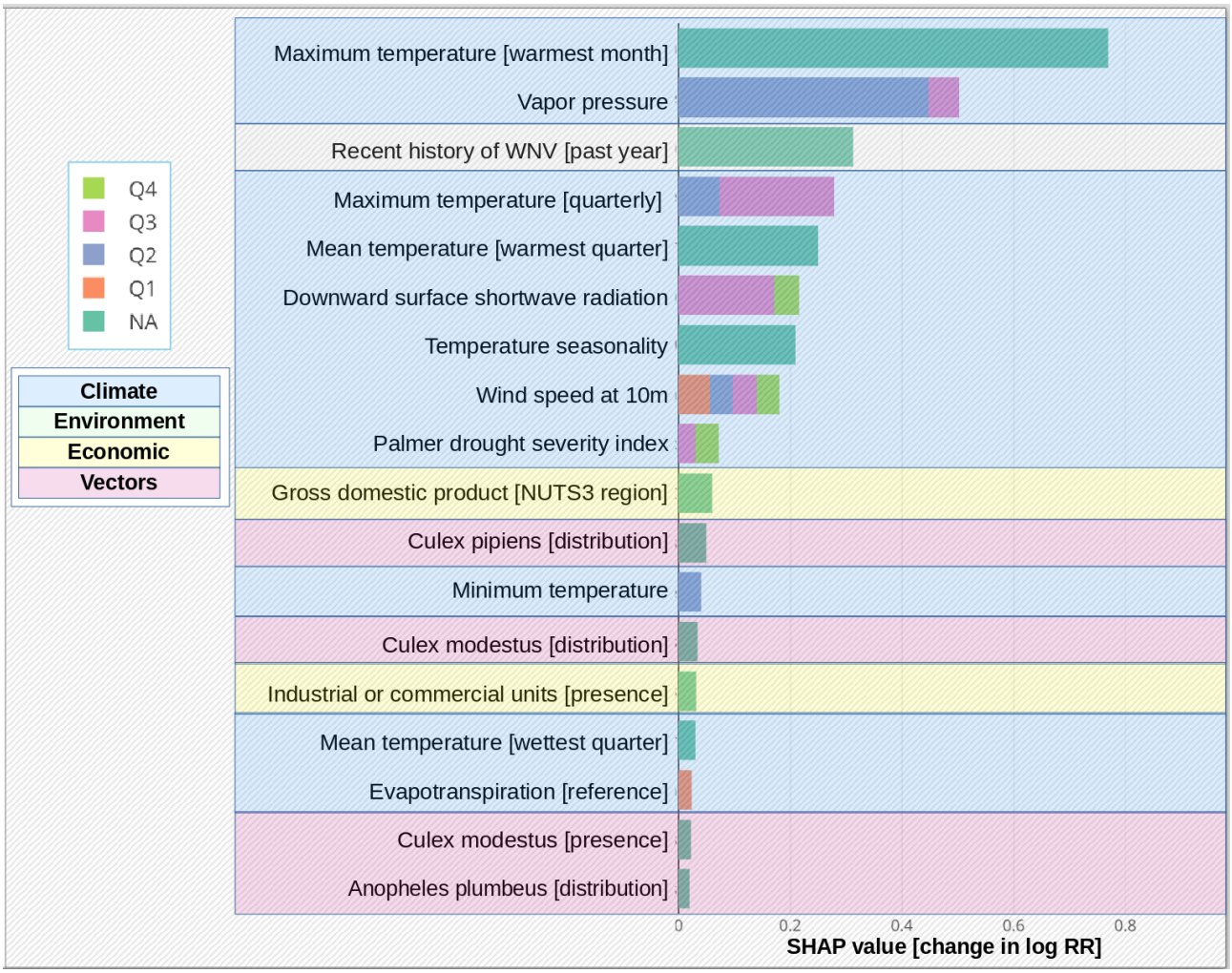
Geospatial disease models are notoriously difficult to compare. All differ in terms of included feature classes, aggregation and summary of features, included covariates, and spatiotemporal coverage and resolution. Studies also differ in terms of disease outcomes being predicted – presence versus incidence – as well as case definition and inclusion criteria. Beyond that, relationships between disease and the various features being tested may vary based on specific geography[1] or analytical resolution and scale[2]. This has been noted most particularly in the case of WNV[3]. For example, specific relationships may exist between climate and disease at broad spatial scales, but not at finer scales where other features may dominate[4]. WNV has long been present in the Old World; the dynamics of emergence and transmission are therefore generally considered more complex than in the Americas[5]. In addition to a broader diversity of potential hosts, vectors, environments, and human population dynamics, the presence of several competing viral lineages makes this indisputable[6]. Few have attempted to tackle this challenge. And none have approached the degree of out-of-sample predictive power demonstrated in the present work, particularly in the context of an extreme event year, such as the 2018 outbreak in Europe. Tran et al (2014)[7] presented a similarly scoped logistic regression model to predict the 2012 and 2013 outbreaks in Europe based on data from the statistically similar 2002-2011 years. Reported AUC was 0.810 and 0.853 (resp.) and the feature set covered was an order of magnitude lower than the present work. The predictors identified as most important were temperature in July, MNDWI in early June, wetlands, trans-Saharan migratory routes, and WNV outbreak in the year prior. Marcantonio et al (2015)[6] predicted incidence in Europe using a linear mixed effects model using a range of predictors broadly equivalent to the present work. Climate, land use, indices of water, vegetation, conservation status, landscape fragmentation and human population density were included. An R^2 of .32 was reported, with climatic and environmental features found to be the most important. Work more similar to the present study has been produced in the American context. Keyel (2019)[8] reported on a similarly high-dimensional model using a similar underlying methodology (Random Forest). An R^2 of .72 was reported, with climate and total human population being the most important features. And similar to the present work, scale-dependent variation of feature effect estimates were noted. Similar conclusions were also drawn regarding the primacy of non-climatic features in determining baseline risk for a given locale. However, predictive power was found to drop off substantially at lower geospatial scales. The present study presents a powerful means to mitigate such concerns. As demonstrated, the "bottom-up" nature of local model aggregation allows for the reporting of standardized effect sizes. This is an extremely common practice in the areas of biostatistics and meta-analyses, and has indeed been cited in those contexts as preferable over direct comparison of effect estimates[9,10]. In such contexts, effect size can be directly converted into a probabilistic measure of the degree to which an observed change is associated with the outcome[11]. It has also been demonstrated to directly correspond with common measures of explanatory power, such a Pearson's r , r^2 , and the non-parametric U (Mann-Whitney) statistic[12]. That this measure also happens to control for scale-dependent variability in the geospatial context is an added plus.

Effects predicted by SHAP align well with the published consensus. Efforts to quantify the effect of various potential determinants on WNV outbreak risk have indeed been well-documented. However, as discussed, broad consensus in the literature with respect to magnitude and direction of effect does not exist. Whatever consensus does exists relies on generalized restatements of the preestablished mechanistic relationships: "ectotherms require heat", "mosquito breeding requires standing water", "trans-Saharan migrant birds carry the disease", "disease is likely to reoccur in previously infected regions", etc[13,14]. However, such generalized relationships have been rarely concretely reflected in effect estimates from any individual model. And even fewer have presented such results side-by-side within the context of a single, comprehensive model. The present work

offers a novel, but intuitive solution. When parameter effects are assessed only for those regions where positive outbreak is indicated (*Figure SS1*), the prevailing consensus emerges. Priority and magnitude of effect of top positive predictors align remarkably well with those reported within the literature[15]. The only notable surprises are with respect to the effects of downward surface shortwave radiation and distribution of the vector, *Anopheles plumbeus*. However, the former aligns well with findings regarding the effect of diurnal variation on vector activity. As for the latter, prior work has confirmed the capacity for *Anopheles plumbeus* to serve as vector for WNV[16] and the ECDC suggests this possibility on their WNV fact sheet (current as of 2020)[17]. Further work is required to quantify the degree of alignment between the these results and those reported in the literature. However, present results demonstrate strong *prima facia* validity.

Figures and Legends

[Figure SS1]



Drivers of positive indication revealed via analysis of aggregate effects. The output of the SHAP reannalysis is an effect matrix that is dimensionally identical to the original data set. Diagnostic analyses (Figure S3) confirmed a high degree of correlative association between the feature effect matrix and the original data. One convenient feature of this surrogate data model is the ability to generate and assess arbitrary aggregations of cases and/or feature sets. We exploited this feature to obtain aggregate effect estimations for the regions where positive outbreak status was indicated by the model in 2018. We found that vapor pressure became a far more dominant predictor in this context as well as the local distribution of the known mosquito vectors, *Culex pipiens* and *Culex modestus*, and the suspected but never confirmed vector, *Anopheles plumbeus*. In this context – that of regions with positive outbreak risk indication in 2018 – the effect of autocorrelative history of outbreak once again became an important determinant of risk.

References

1. Bowden, S. E., Magori, K. & Drake, J. M. Regional Differences in the Association Between Land Cover and West Nile Virus Disease Incidence in Humans in the United States. *Am. J. Trop. Med. Hyg.* **84**, 234–238 (2011).
2. Paull, S. H. *et al.* Drought and immunity determine the intensity of West Nile virus epidemics and climate change impacts. *Proc. R. Soc. B Biol. Sci.* **284**, 20162078 (2017).
3. Petersen, L. R. & Fischer, M. Unpredictable and Difficult to Control — The Adolescence of West Nile Virus. *N. Engl. J. Med.* **367**, 1281–1284 (2012).
4. Cohen, J. M. *et al.* Spatial scale modulates the strength of ecological processes driving disease distributions. *Proc. Natl. Acad. Sci.* **113**, E3359–E3364 (2016).
5. Camp, J. V. & Nowotny, N. The knowns and unknowns of West Nile virus in Europe: what did we learn from the 2018 outbreak? *Expert Rev. Anti Infect. Ther.* **18**, 145–154 (2020).
6. Marcantonio, M. *et al.* Identifying the Environmental Conditions Favouring West Nile Virus Outbreaks in Europe. *PLoS ONE* **10**, (2015).
7. Tran, A. *et al.* Environmental predictors of West Nile fever risk in Europe. *Int. J. Health Geogr.* **13**, 26 (2014).
8. Keyel, A. C. *et al.* 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 (2019).
9. Takeshima, N. *et al.* Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference? *BMC Med. Res. Methodol.* **14**, 30 (2014).
10. White, I. R. & Thomas, J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. *Clin. Trials* **2**, 141–151 (2005).
11. Acion, L., Peterson, J. J., Temple, S. & Arndt, S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat. Med.* **25**, 591–602 (2006).
12. Cohen, J. Statistical power analysis. *Curr. Dir. Psychol. Sci.* **1**, 98–101 (1992).
13. Esser, H. J. *et al.* Risk factors associated with sustained circulation of six zoonotic arboviruses: a systematic review for selection of surveillance sites in non-endemic areas. *Parasit. Vectors* **12**, 265 (2019).
14. Paz, S. Effects of climate change on vector-borne diseases: an updated focus on

West Nile virus in humans. *Emerg. Top. Life Sci.* **3**, 143–152 (2019).

15. Paz, S. & Semenza, J. C. Environmental Drivers of West Nile Fever Epidemiology in Europe and Western Asia—A Review. *Int. J. Environ. Res. Public. Health* **10**, 3543–3562 (2013).

16. Medlock, J. M., Snow, K. R. & Leach, S. Potential transmission of West Nile virus in the British Isles: an ecological review of candidate mosquito bridge vectors. *Med. Vet. Entomol.* **19**, 2–21 (2005).

17. Anopheles plumbeus - Factsheet for experts. *European Centre for Disease Prevention and Control* https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/anopheles-plumbeus.