Put Title Here

Anne Marie Saunders

Imperial College London

MSc Computational Methods in Ecology and Evolution

August 27th, 2020

Declaration

This is the declaration.

Abstract

This is the abstract

Contents

Introduction 1
Methods 4
Mosquito Abundance Data 4
Meteorological Data 5
Data Pre-Processing 5
Model Structure 5
Model Selection and Evaluation 7
Results 8
Discussion 11

Introduction

Introduce

At the end of the 20th century, the world experienced a global resurgence of vector-borne disease (Gubler, 2001). Diseases that had been well-controlled in the early-to-mid 1900s surged under complacent public health policies and insufficient research funding (Gubler, 1998). Today, vector-borne diseases represent 17% of the total global infectious disease burden and cause millions of deaths annually. Mosquito-borne diseases, such as malaria, dengue, Zika, yellow fever, West Nile virus, and chikungunya, singularly infect more than an estimated 200 million individuals worldwide every year (WHO, 2017).

Surveillance of mosquito populations is a successful method to control the public health impacts of vector-borne disease (Vazquez-Prokopec et al., 2010). Intervention in growing populations through chemical control measures can effectively reduce disease incidence (Tomerini et al., 2011). Sampling methods are, however, often limited by resource constraints (Sedda et al., 2019). Many studies have attempted to develop early warning models of disease incidence through prediction, rather than surveillance, of mosquito abundance (Beck-Johnson et al., 2013; Li et al., 2019; Poh et al., 2019). Predictive models of mosquito abundance could offer a cost-effective strategy with which to plan control measures (Yang et al., 2009). In order to understand the potential of applying early warning models of disease incidence to resource-poor sites, greater investigation is needed into optimal methodologies of vector abundance models across locations and species.

Mosquito abundances are affected by many factors, including land-use, elevation, and vegetation cover, but meteorological variables, such as temperature and precipitation, are particularly predictive of population dynamics and commonly used in abundance models (Trawinski and MacKay, 2008; **?**; **?**).

Precipitation could have mixed effects on mosquito abundance dynamics. Precipitation raises near-surface humidity, which increases adult mosquito activity and host-seeking behaviour (Shaman and Day, 2007). This increased reproductive activity would lead to lagged effects on mosquito abundance, but would also cause immediate increases in mosquito trap counts. Many trap types are designed to attract host-seeking or gravid females and are consequently also attractive to mate-seeking male mosquitoes (Li et al., 2016). Rainfall also affects many of the aquatic habitats important for early life stages (Shaman and Day, 2007). Container-breeding species that use man-made containers for oviposition may experience increased breeding sites when precipitation creates habitats in otherwise dry containers (Keith, 2005). Precipitation can also expand suitable habitats for mosquitoes breeding in natural water bodies (Koenraadt and Harrington, 2008). While some rainfall seems likely to have positive effects on mosquito abundance, the effect of heavy rainfall is less clear. Heavy rainfall can flush immature mosquitoes from aquatic habitats, but the extent of this effect varies among species (Koenraadt and Harrington, 2008; Paaijmans et al., 2007). For example, specialist container-breeding *Aedes aegypti* has been found to have a stronger protective diving response to rainfall compared to generalist habitat breeding *Culex pipiens* (Koenraadt and Harrington, 2008).

Cover: - temperature effects - lags - GAMS - non-linear relationship between climate and mosquito abundance (Roiz et al., 2014)

Gaps

- many species/locations - use of autoregressive term or not

- different temporal resolutions (so far restricted) - lag determination across species/location

Novelty of Research

Justify choice of Florida

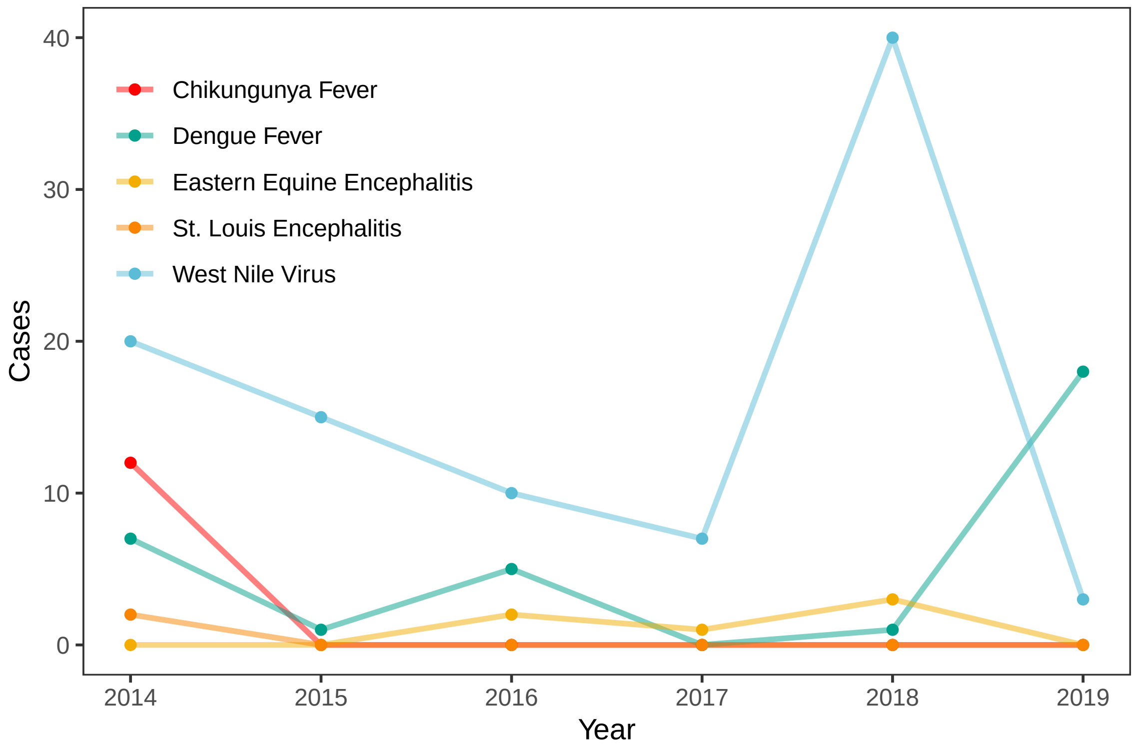


Fig. 1. Number of locally transmitted cases of mosquito-borne disease in humans in Florida between 2014 and 2019, as reported by the Florida Department of Health in 2020. Zika was also locally transmitted in 2016 (300 confirmed cases) and 2017 (2 confirmed cases) but are not shown due to the limited scale of the y axis. Imported cases, where infection was contracted elsewhere but reported in Florida, are not included as these were not contracted from Florida mosquitoes.

More sustainable systems are needed for vector-borne disease surveillance (Vazquez-Prokopec et al., 2010)

Many species, many locations

assessing practical application by comparing predictive power of autoregressive vs non autoregressive model

Clear-cut Questions

1. **Which level of temporal resolution of temperature and precipitation data is best able to predict mosquito abundances?**

2. **To what extent can species-specific abundance models be predictive at multiple locations?**

3. **How does incorporation of an auto-regressive term affect the predictive ability of temperature and precipitation driven models of mosquito abundance?**

Methods

Mosquito Abundance Data

Mosquito count data was obtained from the VectDyn database (**FIXME CITE**). VectDyn is a global database containing spatially and temporally explicit abundance data of mosquitoes and other arthropod vectors from published data and surveillance program records. Mosquito abundance from 205 global locations was narrowed down to seven data-rich counties in Florida, U.S.A. I decided to use data from five of these counties which had multi-year surveillance records and nearly year-round sampling from which fairly continuous time series of mosquito abundances could be formed.

Each location contained multiple trap sites. A variety of trap types were used, including BG-Sentinel traps, CDC light traps, animal-baited traps, and CDC gravid traps. Trap type can affect the efficacy of trapping for different mosquito species (Li et al., 2016). At locations with multiple traps, proportionality of each trap type was inconsistent across the time series. This would complicate conclusions of species abundances aggregated across multiple trap types. Thus, for locations with multiple types of traps, I identified the most common trap type and removed observations from all other trap types.

I averaged species-specific count data in each location and at weekly, biweekly, and monthly temporal resolutions. This transformed integer count values to averaged indicators of county-level abundance and allowed me to account for frequent variation in the number of traps deployed for each location. I aggregated species that are morphologically indistinguishable from one another according to the **ontology used by VectorBase- FIXME CITE/REFERENCE** (**list in SI**). I then removed species that had only zero or NA abundance values. Species counts that were identified only to the genus or family level were also removed. After this processing, I had 163 datasets of unique species and location combinations aggregated at weekly, biweekly, and monthly resolutions). Table 1 summarises the results of this processing at my five focal locations.

County	Trap Type	Years of Data	Number of Species
Lee	CDC Light Trap	2008 - 2017	18
Manatee	CDC Light Trap	2012 - 2016	48
Orange	CDC Light Trap	2012 - 2017	31
Saint Johns	baited light trap	2004 - 2015	38
Walton	New Jersey Trap	2014 - 2017	28

Table 1. Locations of abundance data and associated length of data collection, trap type, and number of species at each location. Individual species in these locations may have abundance records that are shorter than the overall years of data of each location.

Meteorological Data

Temperature and precipitation datasets were obtained from the NOAA Climate Data Online database as global NetCDF raster files at a spatial resolution of 0.50 degrees latitude and 0.50 degrees longitude. Maximum temperature in Celsius and total daily precipitation in millimetres were used based on availability of data. Rasters were rotated 180 degrees to match coordinate rotation of trap locations. Maximum daily temperature and total precipitation values were then extracted by taking the mean of the bilinear interpolation of the 4 closest raster cells to each trap location. Each raster file contained one year of meteorological data, and so this procedure was repeated for each year of the surveillance period. I then mapped the extracted daily maximum temperature and total daily precipitation to corresponding mosquito abundances by date and trap location.

Data Pre-Processing

I spatially and temporally aggregated maximum temperature data by averaging both at county-level and at weekly, biweekly, and monthly time scales. Consequently, temperature in this study refers to the average maximum daily temperature across respective temporal scales. **FIXME explain why choice of max temp vs average, min, etc**

Number of days of rainfall has been shown to be a more effective predictor of mosquito-borne disease incidence than cumulative precipitation (Xu et al., 2017). This is likely due to the maintenance of humid conditions over time with frequent rainfall. Humidity has been independently assessed as a significant predictor of abundance dynamics (Trawinski and MacKay, 2008), and so this representation of precipitation may capture both humidity and precipitation effects. With this in mind, I aggregated precipitation to weekly, biweekly, and monthly scales by summing the number of days in each temporal period with non-zero cumulative precipitation. Consequently precipitation as I will further refer to it refers to a discrete number of days of rainfall, with a maximum of 7, 14, or 31 dependent on temporal scale. Finally, I removed rows of data with missing values in any response or explanatory variables.

Model Structure

I used univariate generalized additive models for each dataset to determine the best-fit temporal lags between abundance and temperature as well as between abundance and precipitation. Because aggregated abundance values are positive, non-integer, and non-normally distributed, I used a Gamma family distribution with a log-link function. These models had the form:

$$ln(V_t + 1) = a_0 + f_1(T_{t-l_T}) + \epsilon_t \quad (1)$$

$$ln(V_t + 1) = a_0 + f_1(P_{t-l_P}) + \epsilon_t \quad (2)$$

V_t is abundance at time t, a_0 is the intercept, T_{t-l_T} and P_{t-l_P} are the temperature and precipitation at l_T and l_P, respectively, time periods prior to time t. One is added to abundance at time t in order to prevent undefined values from the logarithm of zero abundance values. f_1 is a smooth function comprised of cubic polynomial basis functions. Each smooth function was allowed up to nine basis functions to account for flexibility in the model fit. The penalisation process in GAM fitting reduces the number of basis functions to an optimised number for each smooth function. If a smooth function had fewer than ten unique data points, such as with weekly precipitation datasets with values between zero and seven, this upper limit needed to be lowered. The maximum number of basis functions in these cases was set to one less than the number of unique values for that smooth function.

The use of cubic polynomial basis functions is an effective way to avoid the underestimation of standard error caused by concurvity with non-parametric basis functions (Dominici et al., 2002; Ramsay et al., 2003). Concurvity is the extent to which each smooth function can be approximated by other smooth function predictors and is analogous to multicollinearity in linear models. Because of the way variances are estimated in GAMs, concurvity can cause cause underestimation of p-values and thus lead to Type I errors (Ramsay et al., 2003). I used shrinkage methods to allow for full penalisation of spline complexity and REML (restricted maximum likelihood) for optimising smooth parameter estimates. Both of these smooth function estimation methodologies have been shown to maximise the predictive ability of GAMs (Marra and Wood, 2011).

Each dataset was also fit with two multivariate models, each incorporating temperature and precipitation, and one of which incorporating an autoregressive term.

$$ln(V_t + 1) = a_0 + f_1(T_{t-l_T}) + f_2(P_{t-l_P}) + \epsilon_t \quad (3)$$

$$ln(V_t + 1) = a_0 + f_1(T_{t-l_T}) + f_2(P_{t-l_P}) + f_3(V_{t-1} + 1) + \epsilon_t \quad (4)$$

In Equation 4, V_{t-1} is the abundance one time step prior to abundance at time t.

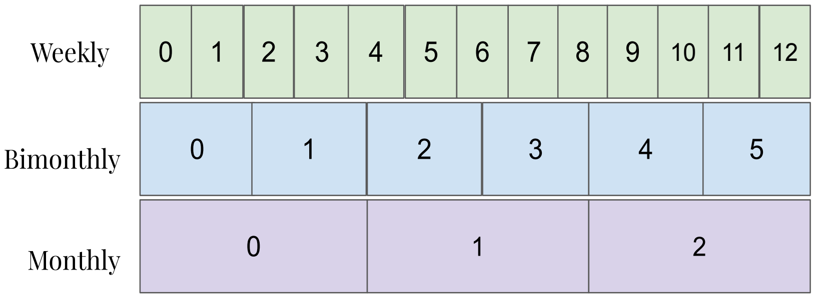


Fig. 2. Lags of temperature (l_T) and precipitation (l_P) considered at each temporal resolution. The numbers in each box represent the values of l_T and l_P. A two month lag, for example, would include data from two to three months prior or, on average, about 61 to 91 days prior. The most equivalent weekly lags (lags 9 through 12) contain data from 63 to 91 days prior. A wide range of lags were tested for model selection in order to avoid making strict assumptions about the relationship of meteorological variables with abundance and to account for wide variation in mosquito development time (Barrera and Medialdea, 1996; Beck-Johnson et al., 2013)

Model Selection and Evaluation

Models 1 and 2 were used to determine the best-fit lags of temperature and precipitation for each dataset at each temporal aggregation. At the weekly and biweekly scale, lags between zero and six weeks were considered. At the monthly scale, lags between zero and two months were considered. Best-fit lags for each temporal scale were chosen by finding the minimum Akaike's Information Criterion (AIC) value among models fit with each possible lag length.

$$AIC = -2ln[L(\hat{\theta}_p|y)] + 2p \tag{5}$$

AIC is a method for determining the likelihood of a given model while penalizing model complexity (Johnson and Omland, 2004).

The Akaike weight of each best fit model was calculated in order to indicate the confidence in this lag as compared to other univariate lagged models.

FIXME- I have just realized that this is an inappropriate use of AIC. These univariate models are not nested and are based on different datasets- different lagged meteorological data. I should actually use GCV- generalized cross validation.

Once best-fit lags of meteorological variables for each dataset and temporal scale were determined, these lags were incorporated into multivariate models 3 and 4. The generalized cross-validation (GCV) score and deviance explained of these multivariate models were compared.

Results

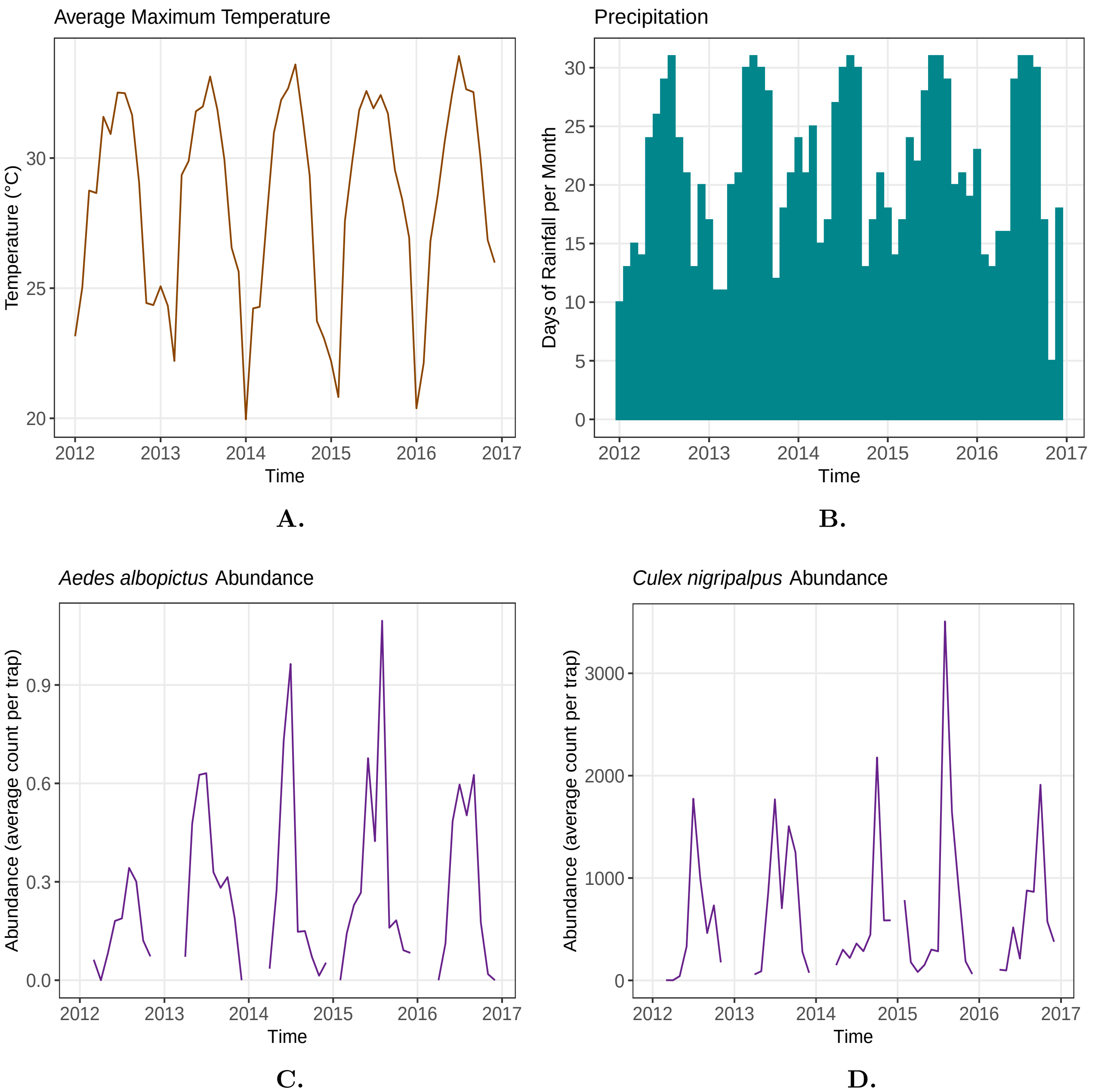


Fig. 3. Time series data of temperature, precipitation, and two vector species from Manatee County, Florida, 2012-2016. Data shown here are aggregated at the weekly scale. In **A.**, average daily maximum temperature per week shows regular seasonality with a range of about 15 °C. In **B.**, precipitation is very frequent during during a summer rainy season. This is common in all locations. **C.** and **D.** show patterns of abundance for two vector species, *Aedes albopictus* and *Culex nigripalpus*. Abundance datasets were non-continuous over winter periods where abundances are assumed to be low so traps are not employed.

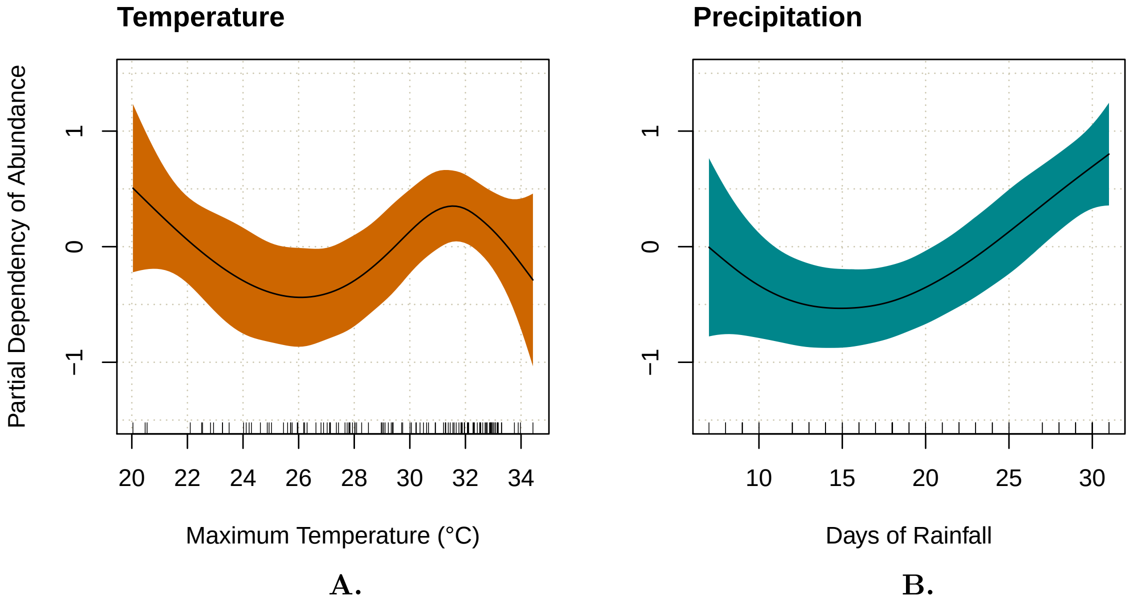


Fig. 4. Sample of multivariate model fitting (Equation 3) showing the partial dependency of *Anopheles quadrimaculatus* abundance in Lee County, Florida on temperature and precipitation (Equation 3) at the monthly temporal aggregation. Both temperature and precipitation are significant and best fit at a lag of two months.

Median Deviance Explained at Various Temporal Resolutions			
County	Weekly	Bimonthly	Monthly
Lee	37.7%	45.8%	*52.2%
Manatee	30.2%	35.5%	*40.1%
Orange	26.2%	29.0%	*35.0%
Saint Johns	13.0%	*25.9%	12.6%
Walton	31.7%	40.2%	*49.9%
All Counties	28.6%	33.0%	*37.9%

Table 2. Deviance explained by the best fit multivariate model of temperature and precipitation for each dataset summarised by the median value at each location and temporal resolution. Median was used because the distribution of deviance explained was left skewed. The best-performing temporal resolution in each row is marked with an asterisk.

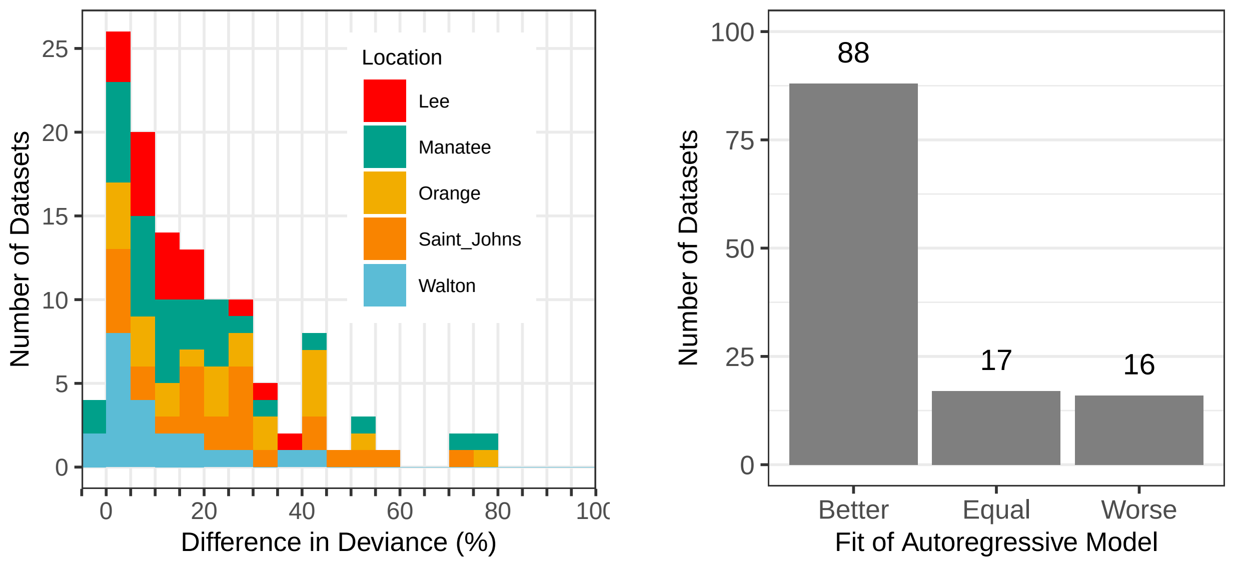
Median MAE at Various Temporal Resolutions			
County	Weekly	Bimonthly	Monthly
Lee	*3.6	9.66	11.87
Manatee	1.04	0.92	*0.90
Orange	0.50	0.44	*0.43
Saint Johns	0.20	0.20	0.20
Walton	0.26	0.22	*0.19
All Counties	*0.76	0.83	0.78

Table 3. This table is an alternative to Table 2. Deviance is a relative measure that summarises the percent of deviance of the dataset that is explained by the model. MAE (Mean Absolute Error), on the other hand, is an absolute measure that can be interpreted as the average error in the mosquitoes predicted for each dataset. It is non-directional. I chose MAE over RMSE because RMSE would be larger for datasets with more data points (weekly level), fit nonwithstanding. This info will be in methods. The real caption will mirror the caption of the alternative Deviance Explained table.

Table 4. Significance of temperature and precipitation in seven major vector species across Florida locations according to a multivariate GAM (Equation 3) constructed with the best fit lags of each variable at a monthly resolution. Each species was present in every location. Significance is marked with by the presence of a T for temperature and P for precipitation. Total locations with significant temperature and precipitation are summed for each species

Locations	*Aedes*[1]		*Aedes*[2]		*Anopheles*[1]		*Culex*[1]		*Culex*[2]		*Culiseta*[1]	
Lee		P		P	T	P	T		T	P	T	P
Manatee	T		T	P	T	P	T		T	P	T	
Orange	T	P	T		T		T		T	P		P
St. Johns	T	P		P			T	P		P		P
Walton	T	P	T	P			T		T		T	P
Total P		4		4		1		1		4		4
Total T	4		3		3		5		4		3	

Aedes[1] = *Aedes atlanticus tormentor morphological group; Aedes*[2] = *Aedes albopictus; Anopheles*[1] = *Anopheles quadrimaculatus; Culex*[1] = *Culex pipiens morphological group; Culex*[2] = *Culex nigripalpus; Culiseta*[1] = *Culiseta melanura*



A. **B.**

Fig. 5. Comparison of the fit of autoregressive versus non-autoregressive meteorological models of mosquito abundance for each monthly dataset (Equations 3 and 4, respectively). After removing 8 datasets where the autoregressive model did not converge, 121 datasets were compared. In **A.** a histogram of the difference in deviance explained between the autoregressive and non-autoregressive model for each dataset is shown. Mean change in deviance explained with the incorporation of an autoregressive term was **+18.7%** across all locations. Four datasets had a decrease (range of [-2.1%, -0.2%]) in deviance explained with incorporation of an autoregressive term. In **B.**, the number of datasets where the autoregressive model was a better, equal, and worse fit than the non-autoregressive model is shown. Models were selected using AIC. AIC differences of less than 2 were considered to be equally best fit (Johnson and Omland, 2004)

Discussion

Mosquito-borne disease transmission is dependent on both natural and human environmental factors, such as climate, land use, health infrastructure, host demographics- using only climate will overestimate impact of climate change on populations (Parham et al., 2015)

Aggregation of precipitation will lose the importance of quick torrential rains likely to flush immature mosquitoes (Koenraadt and Harrington, 2008)

Even if monthly aggregated data captures more of the variation, we're talking aggregation not surveillence protocol. Studies have shown lower sampling error when doing more frequent (1/week) sampling versus less frequent but more sampling (2/fortnight) (Magbity and Lines, 2002)

Bibliography

(2017). Global vector control response 2017–2030. Technical report, World Health Organization, Geneva.

(2020). Mosquito-Borne Disease Surveillance.

Barrera, R. and Medialdea, V. (1996). Development time and resistance to starvation of mosquito larvae. *Journal of Natural History*, 30(3):447–458.

Beck-Johnson, L. M., Nelson, W. A., Paaijmans, K. P., Read, A. F., Thomas, M. B., and Bjørnstad, O. N. (2013). The effect of temperature on Anopheles mosquito population dynamics and the potential for malaria transmission. *PLoS ONE*, 8(11).

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, 156(3):193–203.

Gubler, D. J. (1998). Resurgent vector-borne diseases as a global health problem. *Emerging Infectious Diseases*, 4(3):442–450.

Gubler, D. J. (2001). Prevention and control of tropical diseases in the 21st century: Back to the field. *American Journal of Tropical Medicine and Hygiene*, 65(1).

Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*, 19(2):101–108.

Keith, J. (2005). The importance of agricultural tire habitats for mosquitoes of public health importance in New York State. *Source: Journal of the American Mosquito Control Association*, 21(2):171–176.

Koenraadt, C. and Harrington, L. C. (2008). Flushing Effect of Rain on Container-Inhabiting Mosquitoes Aedes aegypti and Culex pipiens (Diptera: Culicidae) . *Journal of Medical Entomology*, 45(1):28–35.

Li, R., Xu, L., Bjørnstad, O. N., Liu, K., Song, T., Chen, A., Xu, B., Liu, Q., and Stenseth, N. C. (2019). Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *Proceedings of the National Academy of Sciences of the United States of America*, 116(9):3624–3629.

Li, Y., Su, X., Zhou, G., Zhang, H., Puthiyakunnon, S., Shuai, S., Cai, S., Gu, J., Zhou, X., Yan, G., and Chen, X.-G. (2016). Comparative evaluation of the efficiency of the BG-Sentinel trap, CDC light trap and Mosquito-oviposition trap for the surveillance of vector mosquitoes. *Parasites & Vectors*, 9(1):446.

Magbity, E. and Lines, J. (2002). Spatial and temporal distribution of Anopheles gambiae s.l. (Diptera: Culicidae) in two Tanzanian villages: implication for designing mosquito sampling routines. *Bulletin of Entomological Research*, 92(6):483–488.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55(7):2372–2387.

Paaijmans, K. P., Wandago, M. O., Githeko, A. K., and Takken, W. (2007). Unexpected High Losses of Anopheles gambiae Larvae Due to Rainfall. *PLoS ONE*, 2(11):e1146.

Parham, P. E., Waldock, J., Christophides, G. K., Hemming, D., Agusto, F., Evans, K. J., Fefferman, N., Gaff, H., Gumel, A., Ladeau, S., Lenhart, S., Mickens, R. E., Naumova, E. N., Ostfeld, R. S., Ready, P. D., Thomas, M. B., Velasco-Hernandez, J., and Michael, E. (2015). Climate, environmental and socio-economic change: Weighing up the balance in vector-borne disease transmission. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1665):1–17.

Poh, K. C., Chaves, L. F., Reyna-Nava, M., Roberts, C. M., Fredregill, C., Bueno, R., Debboun, M., and Hamer, G. L. (2019). The influence of weather and weather variability on mosquito abundance and infection with West Nile virus in Harris County, Texas, USA. *Science of the Total Environment*, 675:260–272.

Ramsay, T. O., Burnett, R. T., and Krewski, D. (2003). The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter. *Epidemiology*, 14(1):18–23.

Roiz, D., Ruiz, S., Soriguer, R., and Figuerola, J. (2014). Climatic effects on mosquito abundance in Mediterranean wetlands. *Parasites and Vectors*, 7(1):1–13.

Sedda, L., Lucas, E. R., Djogbénou, L. S., Edi, A. V. C., Egyir-Yawson, A., Kabula, B. I., Midega, J., Ochomo, E., Weetman, D., and Donnelly, M. J. (2019). Improved spatial ecological sampling using open data and standardization: an example from malaria mosquito surveillance. *Journal of The Royal Society Interface*, 16(153):20180941.

Shaman, J. and Day, J. F. (2007). Reproductive Phase Locking of Mosquito Populations in Response to Rainfall Frequency. *PLoS ONE*, 2(3):e331.

Tomerini, D. M., Dale, P. E., and Sipe, N. (2011). Does mosquito control have an effect on mosquito-borne disease? the case of ross river virus disease and mosquito management in Queensland, Australia. *Journal of the American Mosquito Control Association*, 27(1):39–44.

Trawinski, P. R. and MacKay, D. S. (2008). Meteorologically conditioned time-series predictions of West Nile virus vector mosquitoes. *Vector-Borne and Zoonotic Diseases*, 8(4):505–521.

Vazquez-Prokopec, G. M., Chaves, L. F., Ritchie, S. A., Davis, J., and Kitron, U. (2010). Unforeseen Costs of Cutting Mosquito Surveillance Budgets. *PLoS Neglected Tropical Diseases*, 4(10):e858.

Xu, L., Stige, L. C., Chan, K. S., Zhou, J., Yang, J., Sang, S., Wang, M., Yang, Z., Yan, Z., Jiang, T., Lu, L., Yue, Y., Liu, X., Lin, H., Xu, J., Liu, Q., and Stenseth, N. C. (2017). Climate variation drives dengue dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 114(1):113–118.

Yang, G. J., Brook, B. W., and Bradshaw, C. J. (2009). Predicting the timing and magnitude of tropical mosquito population peaks for maximizing control efficiency. *PLoS Neglected Tropical Diseases*, 3(2):1–9.