bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A dynamic neural network model for real-time prediction of the Zika epidemic in the Americas

Mahmood Akhtar[1*], Moritz U.G. Kraemer[2,3,4], Lauren M. Gardner[1,5ψ]

[1]School of Civil and Environment Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[2] Department of Zoology, University of Oxford, Oxford OX1 3PS, UK
[3] Computational Epidemiology Group, Boston Children's Hospital, Boston, MA, USA
[4] Harvard Medical School, Boston, MA, USA
[5] Johns Hopkins University, Baltimore, MD, USA

Corresponding email: *m.akhtar@unsw.edu.au
Corresponding email: ψl.gardner@unsw.edu.au

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Background
In 2015 the Zika virus spread from Brazil throughout the Americas, posing an unprecedented challenge to the public health community. During the epidemic, international public health officials lacked reliable predictions of the outbreak's expected geographic scale and prevalence of cases, and were therefore unable to plan and allocate surveillance resources in a timely and effective manner.

Methods
In this work we present a dynamic neural network model to predict the geographic spread of outbreaks in real-time. The modeling framework is flexible in three main dimensions i) selection of the chosen risk indicator, *i.e.*, case counts or incidence rate, ii) risk classification scheme, which defines the relative size of the high risk group, and iii) prediction forecast window (one up to 12 weeks). The proposed model can be applied dynamically throughout the course of an outbreak to identify the regions expected to be at greatest risk in the future.

Results
The model is applied to the recent Zika epidemic in the Americas at a weekly temporal resolution and country spatial resolution, using epidemiological data, passenger air travel volumes, vector habitat suitability, socioeconomic and population data for all affected countries and territories in the Americas. The model performance is quantitatively evaluated based on the predictive accuracy of the model. We show that the model can accurately predict the geographic expansion of Zika in the Americas with the overall average accuracy remaining above 85% even for prediction windows of up to 12 weeks.

Conclusions
Sensitivity analysis illustrated the model performance to be robust across a range of features. Critically, the model performed consistently well at various stages throughout the course of the outbreak, indicating it's potential value at the early stages of an epidemic. The predictive capability was superior for shorter forecast windows, and geographically isolated locations that are predominantly connected via air travel. The highly flexible nature of the proposed modeling framework enables policy makers to develop and plan vector control programs and case surveillance strategies which can be tailored to a range of objectives and resource constraints.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The Zika virus, which is primarily transmitted through the bite of infected *Aedes aegypti* mosquitoes [1], was first discovered in Uganda in 1947 [2] from where it spread to Asia in 1960s, where it since has caused small outbreaks. In 2007 ZIKV caused an island wide outbreak in Yap Island, Micronesia [3], followed by outbreaks in French Polynesia [4] and other Pacific islands between 2013−2014 where attack rates where up to 70% [5-7]. It reached Latin America between late 2013 and early 2014, but was not detected by public health authorities until May 2015 [8] and since affected 48 countries and territories in the Americas [9-11]. Since there is no vaccination or treatment available for Zika infections [12, 13], the control of *Ae. aegypti* mosquito populations remains the most important intervention to contain the spread of the virus [14]. In order to optimally allocate resources to suppress vector populations, it is critical to accurately anticipate the occurrence and arrival time of arboviral infections to detect local transmission [15].

Whereas for dengue, the most common arbovirus infection, prediction has attracted wide attention from researchers employing statistical modelling and machine learning methods to guide vector control [16-21], such real-time models do not yet exist for Zika virus. Early warning systems for Thailand, Indonesia, Ecuador and Pakistan have been introduced and are currently in use [22-26]. In addition to conventional predictions based on epidemiological and meteorological data [20, 27, 28], more recent models have successfully incorporated search engines [29, 30], land use [31], human mobility information [32, 33] and spatial dynamics [34-36], and various combinations of the above [37] to improve predictions. Whereas local spread may be mediated by overland travel, continent wide spread is mostly driven by air passenger travel between climatically synchronous regions [38-44].

The aims of our work are to 1) present recurrent neural networks for time ahead predictive modelling as a highly flexible tool for outbreak prediction, and 2) implement and evaluate the model performance for the Zika epidemic in the Americas. The application of neural networks for epidemic risk forecasting has previously been applied to dengue forecasting and risk classification [45-50], detection of mosquito presence [51], temporal modeling of the oviposition of *Aedes aegypti* mosquito [52], *Aedes* larva identification [53], and epidemiologic time-series modeling through fusion of neural networks, fuzzy systems and genetic algorithms [54]. Recently, Jian *et al* [55] performed a comparison of different machine learning models to map the probability of Zika epidemic outbreak using publically available global Zika case data and other known covariates of transmission risk. Their study provides valuable insight into the potential role of machine learning models for understanding Zika transmission; however, it is static in nature, *i.e.*, it does not account for time-series data, and did not account for human mobility, both of which are incorporated in our modelling framework.

Here, we apply a dynamic neural network model for N-week ahead prediction for the 2015-2016 Zika epidemic in the Americas. The model implemented in this work relies on multi-dimensional time-series data at the country (or territory) level, specifically epidemiological data, passenger air travel volumes, vector habitat suitability for the primary spreading vector *Ae. aegypti*, socioeconomic and population data. The modeling framework is flexible in three main dimensions: 1) the preferred risk indictor can be chosen by the policy maker, *e.g.*, we consider outbreak

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

size and incidence rate as two primary indicators of risk for a region, 2) five risk classification schemes are defined, where each classification scheme varies in the threshold used to determine the set of countries deemed "high risk", and 3) it can be applied for a range of forecast windows (1 – 12 weeks). Model performance and robustness is evaluated for various combinations of risk indicator, risk classification level, and forecasting windows. Thus, our work represents the first flexible framework of neural networks for epidemic risk forecasting, that allows policy makers to evaluate and weigh the trade-off in prediction accuracy between forecast window and risk classification schemes. Given the availability of the necessary data, the modelling framework proposed here can be applied in real time to future outbreaks of Zika, and other similar vector-borne outbreaks.

Materials and Methods

Data

The model relies on socioeconomic, population, epidemiological, travel and mosquito vector suitability data. All data is aggregated to the country level, and provided for all countries and territories in the Americas. Each data set and corresponding processing is described in detail below, and summarized in Table 1. The data is available as a supplementary file.

Epidemiological Data

Weekly Zika infected cases for each country and territory in the Americas were extracted from the Pan American Health Organization (PAHO) [57], as described in previous studies [40, 43] (data available: github.com/andersen-lab/Zika-cases-PAHO). Although Zika cases in Brazil were reported as early as May 2015, no case data is available for all of 2015 from PAHO because the Brazil Ministry of Health did not declare the Zika cases and associated neurological and congenital syndrome as notifiable conditions until 17 February of 2016 [57]. The missing numbers of cases from July to December 2015 for Brazil were estimated based on the positive correlation between *Ae. aegypti* abundance (described below) and reported case counts as has been done previously [42, 43]. We used smoothing spline [56] to estimate weekly case counts from the monthly reported counts. The weekly country level case counts (Figure 1A) were divided by the total population / 100,000, as previously described [43], to compute weekly *incidence rates* (Figure 1C).

Travel Data

Calibrated monthly passenger travel volumes for each airport-to-airport route in the world were provided by the International Air Transport Associate (IATA) [67], as previously used in [43, 58]. The data includes origin, destination and stopover airport paths for 84% of global air traffic, and includes over 240 airlines and 3,400 airports. The airport level travel was aggregated to a regional level, to compute monthly movements between all countries and territories in the Americas. The *incoming* and *outgoing travel volumes* for each country and territory, originally available from IATA at a monthly temporal resolution, were curve fitted, again using smoothing spline method [56] to obtain corresponding weekly volumes to match with the temporal resolution of our model. In this study, data and estimates from 2015 were also used for 2016, as was done previously [43, 58, 59].

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mosquito Suitability Data

The monthly vector suitability data sets were based on habitat suitability for the principal Zika virus species *Ae. aegypti*, previously used in [43], and initially estimated using original high resolution maps [60] and then enriched to account for seasonal variation in the geographical distribution of *Ae. aegypti* by using time-varying covariate such as temperature persistence, relative humidity, and precipitation as well as static covariates such as urban versus rural areas. The monthly data was translated into weekly data using a smoothing spline [56].

Socioeconomic and Human Population Data

For a country, to prevent or manage an outbreak depends on their ability to implement a successful surveillance and vector control programs [68]. Due to a lack of global data to quantify vector control at country level, we utilized alternative economic and health related country indicators which have previously been revealed to be critical risk factors for Zika spread [43]. A country's economic development can be measured by the *gross domestic product (GDP)* per capita at purchasing power parity (PPP), in international dollars. The figures from World Bank [61] and the U.S. Bureau of Economic Analysis [62] were used to collect GDP data for each country. The number of *physicians* and the number of hospital *beds* per 10,000 people were used to indicate the availability of health infrastructure in each country. These figures for U.S. and other regions in the Americas were obtained from the Centre of Disease Control and Prevention (CDC) [63], WHO World Health Statistics report [64], and the PAHO [65]. Finally, the human *population densities* (people per sq. km of land area) for each region were collected from World Bank [66] and the U.S. Bureau of Economic Analysis [62].

Connectivity-risk Variables

In addition to the raw input variables, novel connectivity-risk variables are defined and computed for inclusion in the model. These variables are intended to capture the risk posed by potentially infected travelers arriving at a given destination at a given point in time, and in doing so, explicitly capture the dynamic and heterogeneity of the air-traffic network in combination with real-time outbreak status. Two variables are chosen, hereafter referred to as *case-weighted travel risk* and *incidence-weighted travel risk*, as defined in equations (1.a) and (1.b), respectively.

$$CR_j^t = \sum_i C_i^t . V_{i,j}^t \qquad \forall t, \forall j, i \neq j \qquad (1.a)$$

$$IR_j^t = \sum_i I_i^t . V_{i,j}^t \qquad \forall t, \forall j, i \neq j \qquad (1.b)$$

For each region *j* at time *t*, CR_j^t and IR_j^t are computed as the sum of product between passenger volume traveling from origin *i* into destination *j* at time *t* ($V_{i,j}^t$) and the state of the outbreak at origin *i* at time t, namely reported cases, C_i^t, or reported incidence rate, I_i^t. Each of these two variables is computed for all 53 countries or territories for each of the 78 epidemiological weeks. The dynamic variables are illustrated in Figure 1, below the raw case counts and incidence rates.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Summary of input data

Description	Original Temporal Resolution	Spatial Resolution	Temporal Disaggregation	Reference
Zika cases (2015)	monthly	country or territory level	estimation, smoothing spline curve fitting	[42, 43, 56]
Zika cases (2016)	weekly	country or territory level	-	Pan American Health Organization (PAHO) [57]
incidence rates	weekly	country or territory level	(case counts/(population / 100,000)	[43]
incoming and outgoing travel volumes (2015)	monthly	country or territory level	smoothing spline curve fitting	International Air Transport Associate (IATA), [56]
incoming and outgoing travel volumes (2016)	monthly	country or territory level	estimation, smoothing spline curve fitting	as previously done [43, 58, 59], [56]
Aedes vector suitability	monthly	country or territory level	smoothing spline curve fitting	[43, 56, 60]
gross domestic product (GDP) per capita	annual	country or territory level	-	World Bank [61], and U.S. Bureau of Economic Analysis [62]
physicians per 1000 people	annual	country or territory level	-	Centre of Disease Control and Prevention (CDC) [63], WHO World Health Statistics report [64], and the PAHO [65]
beds per 1000 people	annual	country or territory level	-	
population densities (people per sq. km of land area)	annual	country or territory level	-	World Bank [66], and the U.S. Bureau of Economic Analysis [62]

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

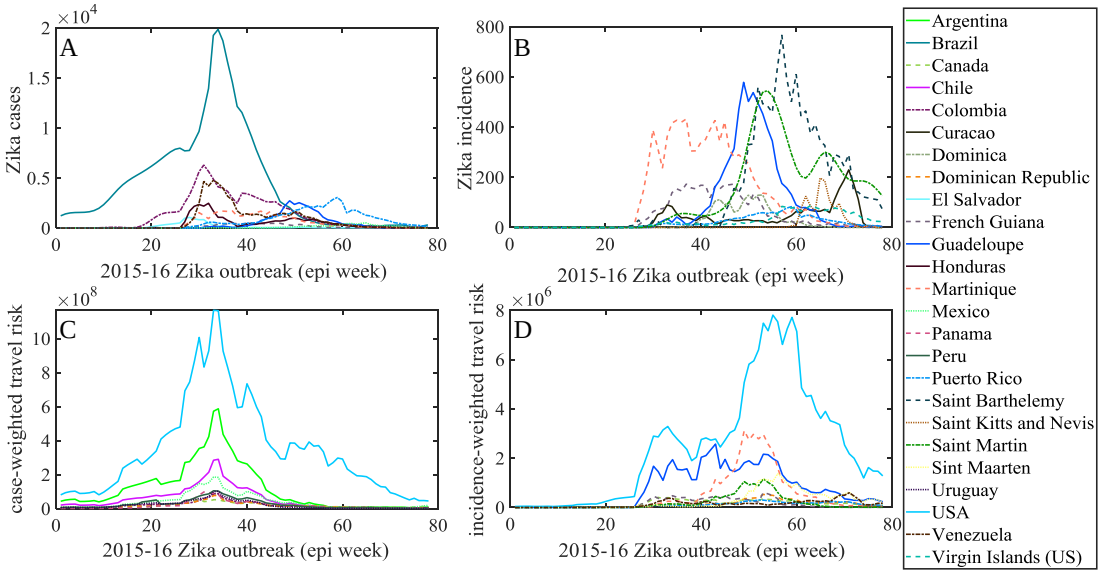


Fig 1. Weekly distribution of case-related input variables. (A) Zika cases and (B) incidence rates in the Americas along with connectivity-risk variables, (C) case-weighted travel risk $\mathbf{CR_j^t}$, and (D) incidence weighted travel risk $\mathbf{IR_j^t}$, for top 10 countries and territories in the Americas.

Neural Network Model

A class of neural architectures based upon Nonlinear AutoRegressive models with eXogenous inputs (NARX) known as NARX neural networks [69-71] is employed herein due to its suitability for modeling of a range of nonlinear systems and computational capabilities equivalent to Turning machines [72]. The NARX networks, as compared to other recurrent neural network architectures, require limited feedback (i.e., feedback from the output neuron rather than from hidden states) and converge much faster with a better generalization [72, 73]. The NARX model can be formalized as follows [72]:

$$y(t) = f\left(x(t), x(t-1), \dots, x(t-d_x); y(t-1), \dots, y(t-d_y)\right) \qquad (2)$$

where $x(t)$ and $y(t)$ denote, respectively, the input and output (or target that should be predicted) of the model at discrete time *t*, while d_x and d_y (with $d_x \geq 1$, $d_y \geq 1$, and $d_x \leq d_y$) are input and output delays called memory orders (Fig. 2). In this work, a NARX model is implemented to provide *N*-step ahead prediction of a time series, as defined below:

$$y_k(t+N) = f\begin{pmatrix} \mathbf{x_1}(t), \mathbf{x_1}(t-1), \dots, \mathbf{x_1}(t-d_x), \dots, \mathbf{x_M}(t), \mathbf{x_M}(t-1), \dots, \mathbf{x_M}(t-d_x), \\ y_k(t), y_k(t-1), \dots, y_k(t-d_y) \end{pmatrix} \qquad (3)$$

Here, $y_k(t+N)$ is the risk classification predicted for the kth region *N* weeks ahead (of present time *t*), which is estimated as a function of $\mathbf{x_m}(t)$ inputs from all $m = 1, 2, \dots, M$ regions for d_x previous weeks, and the previous risk classification state, $y_k(t)$ for region *k* for d_y previous weeks. The prediction model is applied at time *t*, to predict for time *t*+*N*, and therefore relies on data available up until week *t*. That is,

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to predict outbreak risk for epidemiological week X, *N*-weeks ahead, the model is trained and tested using data available up until week (X – N). For example, 12-week ahead prediction for Epi week 40, is performed using data available up to week 28. The function $f(\cdot)$ is an unknown nonlinear mapping function that is approximated by a Multilayer Perceptron (MLP) to form the NARX recurrent neural network [70, 71]. In this work, series-parallel NARX neural network architecture is implemented in Matlab R2018a (The MathWorks, Inc., Natick, Massachusetts, United States) [74].

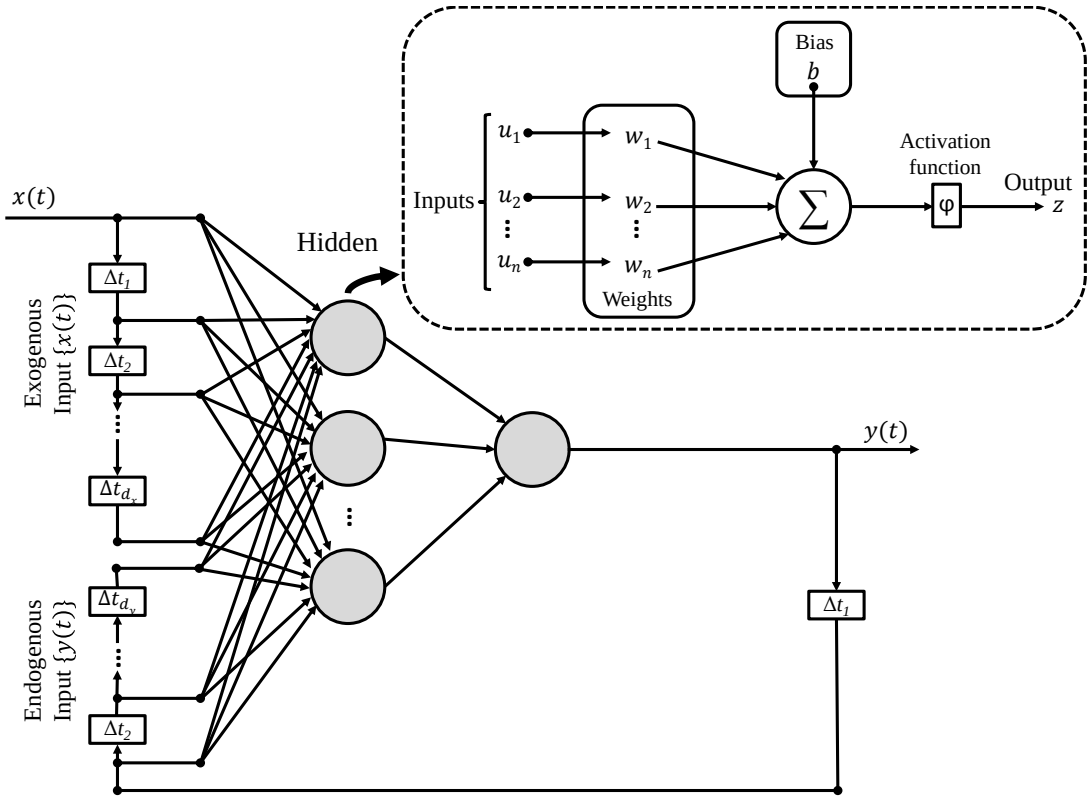


Fig 2. Schematic of NARX network with $\mathbf{d_x}$ input and $\mathbf{d_y}$ output delays: Each neuron produces a single output based on several real-valued inputs to that neuron by forming a linear combination using its input weights and sometimes passing the output through a nonlinear activation function: $\mathbf{z} = \boldsymbol{\varphi}(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i u_i} + \mathbf{b}) = \boldsymbol{\varphi}(\mathbf{w^T x} + \mathbf{b})$, where **w** denotes the vector of weights, **u** is the vector of inputs, **b** is the bias and **φ** is a linear or nonlinear activation function (e.g., Linear, Sigmoid, and Hyperbolic tangent [75]).

In the context of this work, the desired output, $y_k(t+N)$, is a binary risk classifier, *i.e.*, classifying a region *k* as high or low risk at time at time *t*+*N*, for each region, *k*, *N* weeks ahead (of *t*). The vector of input variables for region m at time t is $\mathbf{x}_{\boldsymbol{m}}(t)$, and includes both static and dynamic variables. We consider various thresholds to define the "high risk" group, ranging uniformly between 10% and 50%, where the 10% scheme classifies the 10% of countries reporting the highest number of cases (or highest incidence rate) as high risk, and the other 90% as low risk, similar to [37]. Each thresholds corresponds to a risk classification scheme (*i.e.*, R=10, R=20, etc). Critically, our prediction approach differs from [37], in that our model is trained to predict the risk level directly, rather than predict the number of cases, which are post-processed into risk categories. The performance of the model is evaluated by

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

comparing the estimated risk level (high or low) to the actual risk level for all locations at a specified time, The actual risk level is simply defined at each time period *t* during the outbreak by ranking the regions based on to the number of reported case counts (or incidence rates), and grouping them into high and low risk groups according to the specified threshold.

The static variables used in the model include GDP PPP, population density, number of physicians, and number of hospital beds for each region. The dynamic variables include mosquito vector suitability, outbreak status (both reported case counts and reported incidence rates), total incoming travel volume, total outgoing travel volume, and the two connectivity-risk variables defined as in Equations (1.a) & (1.b), again for each region. Before applying to the NARX model, all data values are normalized to the range [0, 1].

A major contribution of this work is the flexible nature of the model, which allows policy makers to be more or less risk averse in their planning and decision making. Firstly, the risk indicator (used to rank the regions and identify the high risk group) can be chosen by the modeler; in this work we consider two regional risk indicators, i) the number of reported cases and ii) incidence rate. Second, we consider a range of risk classification schemes, which vary by the relative size of the "high risk" group, *i.e.*, R=10, R=20, R=30, R=40, R=50. Third, the forecast window, *N*, is defined to range from *N* = 1, 2, 4, 8 and 12 weeks. Subsequently, any combination of risk indicator, risk classification scheme and forecasting window can be modelled.

In initial settings of the series-parallel NARX neural network, a variety numbers of hidden layer neurons and numbers of tapped delay lines (Eq. (2)) were explored for training and testing of the model. Sensitivity analysis revealed minimal difference in performance of the model under different settings. Therefore, for all experiments presented in this work, the numbers of neural network hidden layer neurons and tapped delay lines are kept constant as two and four, respectively.

To train and test the model, the actual risk classification for each region at each week during the epidemic, $y_k(t)$, was used. For each model run, e.g., a specified risk indicator, risk classification scheme and forecasting window, the input and target vectors are randomly divided into three sets:

1. 70% for training, to tune model parameters minimizing the mean square error between the outputs and targets,
2. 15% for validation, to measure network generalization and to prevent overfitting, by halting training when generalization stops improving (i.e., mean square error of validation samples starts increasing), and
3. 15% for testing, to provide an independent measure of network performance during and after training.

The performance of the model is measured using two metrics: 1) prediction accuracy (ACC) and 2) receiver operating characteristic (ROC) curves. Prediction accuracy is defined as ACC = (TP + TN) / (TP + FP + TN + FN), where true positive (TP) is the number of high risk locations correctly predicted as high risk, false negative (FN) is the number of high risk locations incorrectly predicted as low risk, true negative (TN) is the number of low risk locations correctly predicted as low risk, and false positive

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(FP) is the number of low risk locations incorrectly predicted as high risk. The second performance metric, ROC curve, explores the effects on TP and FP as the position of an arbitrary decision threshold is varied, which in the context of this prediction problem distinguished low and high risk locations. ROC curves were originally developed in 1950s as a technique for visualizing, organizing and selecting classifiers based on their performance [76]. The ROC curve can be characterized as a single number using the area under the ROC curve (AUC), with larger areas having an AUC that approaches 1 indicating a more accurate detection method. In addition to quantifying model performance using these two metrics, we evaluate the robustness of the predictions by comparing the ACC across multiple runs that vary in their selection of testing and training sets (resulting from the randomized sampling). Due to computation time, the robustness is only evaluated for the 4-week forecast window.

Results and Discussion

The model outcome reveals the set of locations expected to be at highest risk at a specified date in the future, *i.e.*, *N* weeks ahead of when the prediction is made. We apply the model for all epidemiological weeks throughout the epidemic, and evaluate performance under each combination of i) risk indicator, ii) classification scheme, and iii) forecast window. For each model run, both ACC and ROC AUC are computed. Results are presented in this section as follows:

1. Country-level Outbreak Prediction.
    a. Performance sensitivity to classification scheme is presented at the country level, for a fixed forecast window (N=4).
    b. Performance sensitivity to forecast window is presented at the country level, for a fixed classification scheme (R=20).
2. Model performance
    a. ACC (averaged over all locations and all EPI weeks) is presented for each classification scheme (i.e., R = 10, 20, 30, 40 and 50) and each forecast window (i.e., *N* = 1, 2, 4, 8 and 12) combination.
    b. ROC AUC is presented for a fixed classification scheme (R=40) and all forecast windows (i.e., *N* = 1, 2, 4, 8 and 12).
    c. Performance sensitivity to epidemiological week is presented for each risk indicator. Results are shown for each classification scheme, and a fixed forecast window (N=4).
    d. Performance (ACC) aggregated by geographic region (*Caribbean*, *South America* and *Central America)* is presented for each classification scheme (i.e., R = 10, 20, 30, 40 and 50) and forecast window (i.e., *N* = 1, 2, 4, 8 and 12).

Country-level Outbreak Prediction

Fig 3. and 4. exemplify the output of the proposed model. Fig 3 illustrates the model predictions at a country-level for a *4*-week prediction window, specifically for Epi week 40, *i.e.*, using data available up until week 36. Fig. 3(A) illustrates the actual risk percentile each country is assigned to in week 40, based on reported case counts. The results presented in the remaining panels of Fig 3 reveal the risk level (high or low) predicted for each country under the five risk classification schemes, namely (B) R=10, (C) R=20, (D) R=30, (E) R=40, and (F) R=50, and whether or not it was correct. For Panels (B)-(E), green indicates a correctly predicted low risk country (TN), light grey indicates an incorrectly predicted high risk country (FP), dark grey

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

indicates an incorrectly predicted low risk country (FN), and the remaining color indicates a correctly predicted high risk country (TP). The inset highlights the results for the *Caribbean* islands. The figure also presents the average ACC over all regions and ACC for just the Caribbean region (grouped similar to [10]) for each classification scheme. For all cases, the predictive capability of the model is similar for the Caribbean as for the entire Americas, and the ACC remains above 90% for R < 30, indicating superior model performance. For example, at Epi week 40, R = 30 and N=4 (using outbreak data and other model variables up to Epi week 36), there were 16 total regions classified as HIGH risk, of which the model correctly identified 13. Furthermore, of the 16 high risk regions, 8 were in the *Caribbean* (i.e., Aruba, Curacao, Dominican Republic, Guadeloupe, Haiti, Jamaica, Martinique, and Puerto Rico), of which the model correctly identified 7. Aruba in the only Caribbean, and Honduras and Panama were the only regions incorrectly predicted as low risk in this scenario. Accurately classifying low risk regions is also important (and assuring the model is not too risk averse). For the same scenario, Epi week 40, R = 30 and N=4, all 18 low risk *Caribbean* locations and 17 of the 19 low risk non-*Caribbean* locations were accurately classified by the model. Paraguay and Suriname were the only regions incorrectly predicted as high risk. These results are consistent with the high reported accuracy of the model, i.e., Overall ACC = 90.15%; *Caribbean* ACC = 96.15%.

Fig 4 illustrates the model predictions at a country-level for varying prediction windows, and a fixed classification scheme of R=20, for Epi week 40. Fig. 4(A) illustrates the actual risk classification (high or low) each country is assigned to in Epi week 40, based on reported case counts. The results presented in the remaining panels of Fig 4 reveal the risk level (high or low) predicted for each country under the five forecasting windows, specifically (B) N=1, (C) N=2, (D) N=4, (E) N=8, and (F) N=12, and whether or not it was correct. For Panels (B)-(E), red indicates a correctly predicted high risk country (TP), green indicates a correctly predicted low risk country (TN), light grey indicates an incorrectly predicted high risk country (FP), dark grey indicates an incorrectly predicted low risk country (FN). The inset highlights the results for the *Caribbean* islands. Similar to Fig. 3, for each forecast window, the reported ACC is averaged both over all regions and for just the Caribbean.

The results reveal that the performance of model, expectedly, deteriorates as the forecast window increases; however, the average accuracy remains above 80% for prediction up to *8*-weeks ahead, and well about 90% for up to 4-weeks ahead. The prediction accuracy for the Caribbean slightly lags the average performance in the Americas. Specifically, for R=20, 5 of the 11 *Caribbean* regions were designated as HIGH risk locations at Epi week 40, *i.e.*, Dominican Republic, Guadeloupe, Jamaica, Martinique, Puerto Rico. For a one-week prediction window, N=1, the model was able to correctly predict 3 of the high risk regions (*i.e.*, Jamaica, Martinique, Puerto Rico), for N=2 it correctly identified two (*i.e.*, Martinique, Puerto Rico), and for N=4, it again correctly identified three (*i.e.*, Guadeloupe, Martinique, Puerto Rico). However, the model did not correctly predict any high risk locations in the Caribbean at N=8 and N=12 window lengths. This error is likely due to the low and sporadic reporting of Zika cases in the region around week 30. Similar prediction capability is illustrated for R=50 (not shown in the figure), in which case out of the 13 *Caribbean* HIGH risk locations, the model correctly identifies all locations at N=1, 2 and 4, 10 of the 13 locations at N=8, and only 1 of the 13 at N=12.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

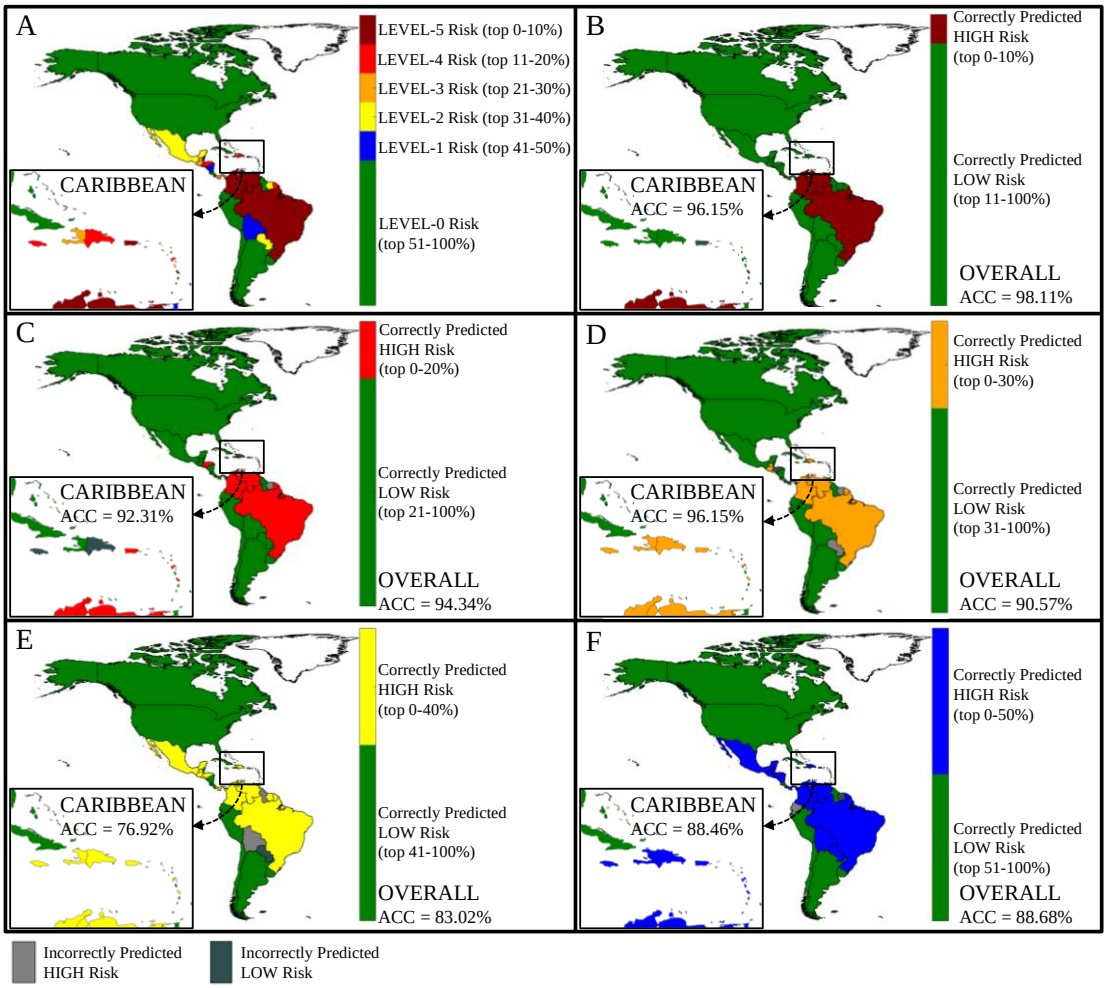


Fig 3. Country prediction accuracy by risk level. Panel (A) illustrates the actual risk level assigned to each country at Epi week 40 for a fixed forecast window, N=4. Panels (B)-(E) each corresponds to a different classification scheme, specifically (B) R=10, (C) R=20, (D) R=30, (E) R=40, and (F) R=50. The inset shown by small rectangle highlights the actual and predicted risk in Caribbean islands. For Panels (B)-(E), green indicates a correctly predicted low risk country, light grey indicates an incorrectly predicted high risk country, and dark grey indicates an incorrectly predicted low risk country. The risk indicator used is case counts.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

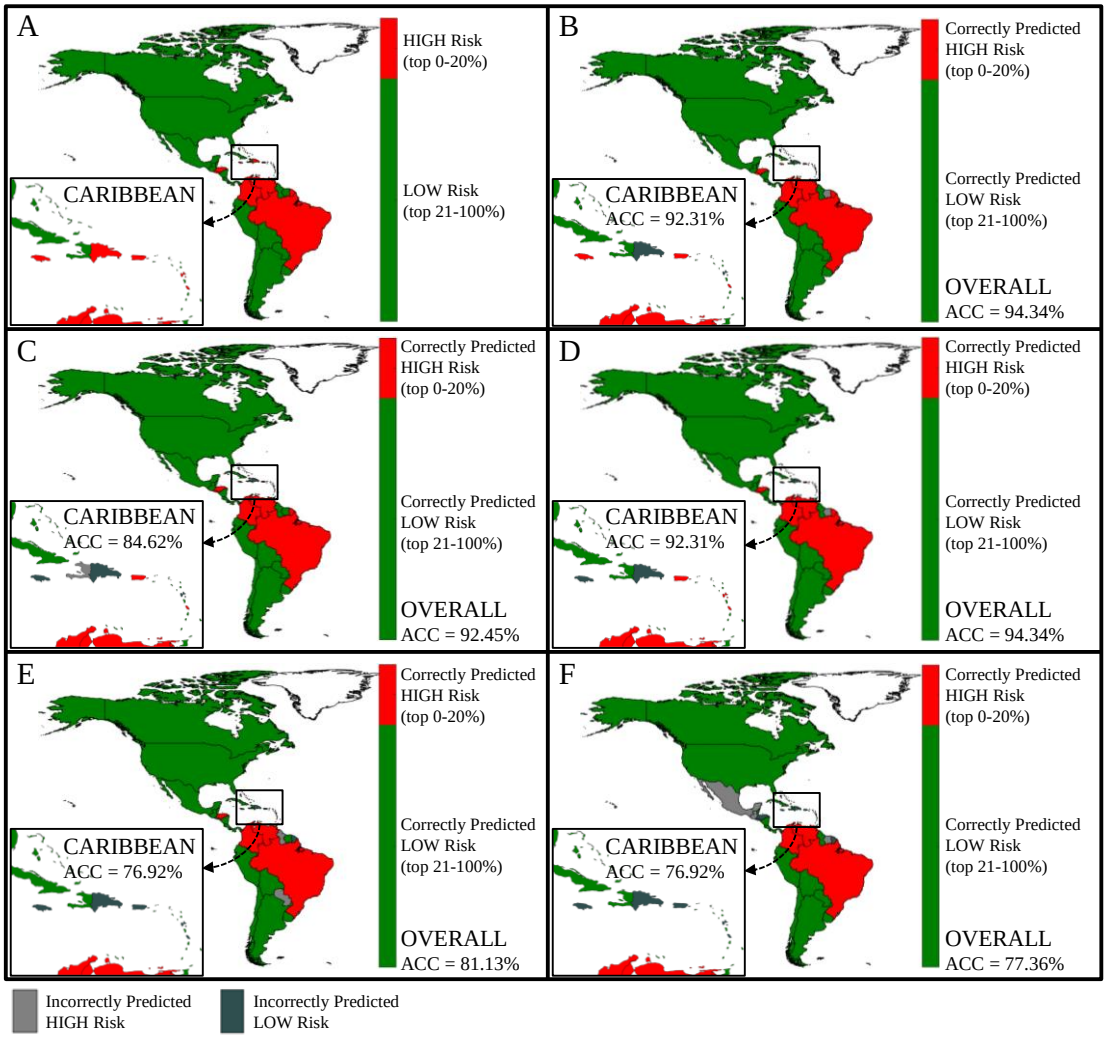


Fig 4. Country prediction accuracy by forecast window. Panel (A) illustrates the actual risk level assigned to each country at Epi week 40 for a fixed classification scheme, R=20. Panels (B)-(E) each corresponds to different forecast windows, specifically (B) N=1, (C) N=2, (D) N=4, (E) N=8, and (F) N=12. The inset shown by small rectangle highlights the actual and predicted risk in Caribbean islands. For Panels (B)-(E), the red indicates a correctly predicted high-risk country and green indicates a correctly predicted low risk country. Light grey indicates an incorrectly predicted high risk country, and dark grey indicates an incorrectly predicted low risk country. The risk indicator used is case counts.

Model performance

The remainder of this section demonstrates the model's performance sensitivity to the range of flexible input parameters available. Fig 5 and 6 illustrate the model performance as a function of classification scheme and forecast window, aggregated over space and time. Specifically, Fig 5 shows the model performance based on ACC, averaged over all locations and all EPI weeks for each combination of risk classification scheme (i.e., R = 10, 20, 30, 40 and 50) and forecast window (i.e., *N* = 1, 2, 4, 8 and 12) (Fig. 5). In general, the performance of the model decreases as the prediction window increases, and as the size of the high risk group increases. When the objective is to identify the top 10% of at-risk regions, the average accuracy of the model remains above 87% for prediction up to *12*-weeks in advance. Further, the model is almost 80% accurate for 4-week ahead prediction for all classification

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

schemes, and almost 90% accurate for all 2-week ahead prediction scenarios, *i.e.*, the correct risk category of 9 out of 10 locations can always be predicted. These results reveal the trade-off between desired forecast window and precision of the high risk group. The quantifiable trade-off between the two model inputs (classification scheme, R, and forecast window, N) can be useful for policies which may vary in desired planning objectives.

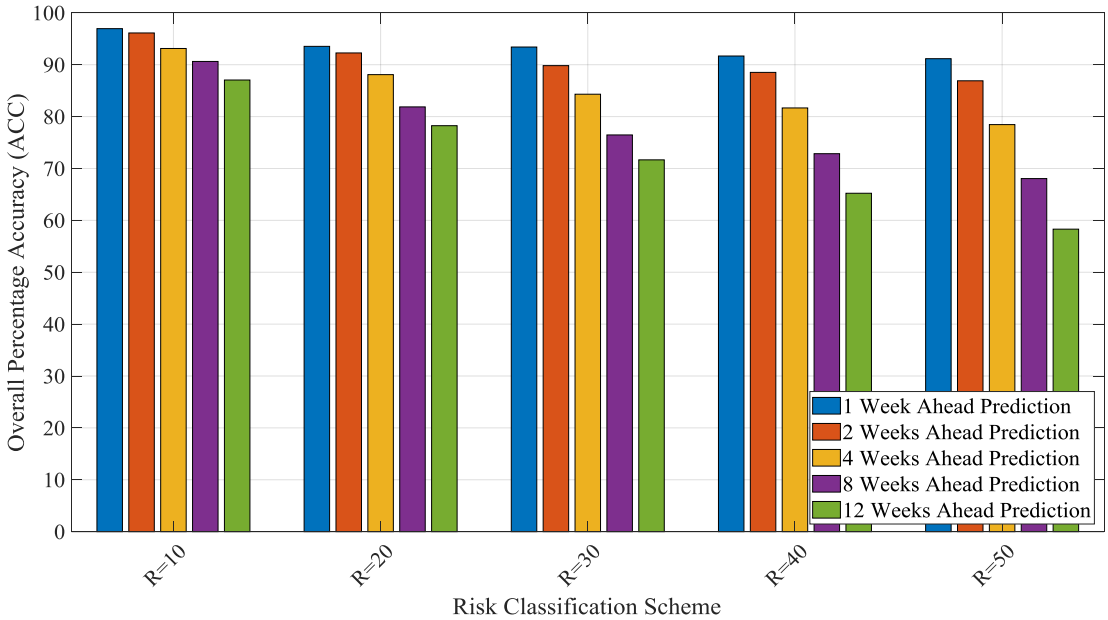


Fig 5. **Aggregate model performance measured by ACC**, (averaged over all locations and all weeks) for all combinations of classification schemes (i.e., R = 10, 20, 30, 40 and 50) and forecast windows (i.e., *N* = 1, 2, 4, 8 and 12), where the risk indicator is case counts.

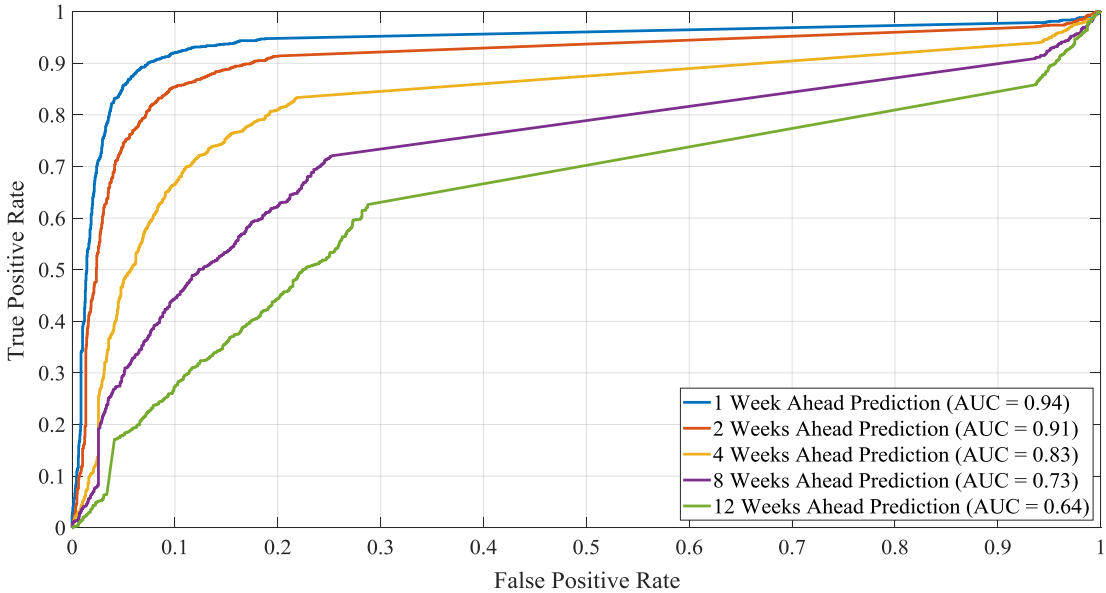


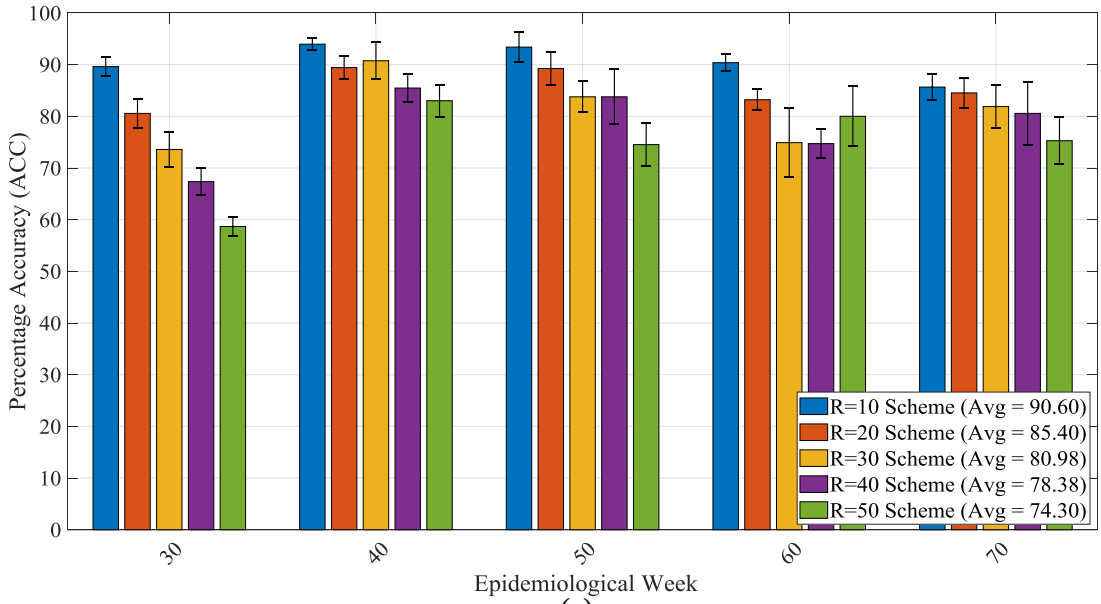
Fig 6. Aggregate model performance measured by ROC AUC (averaged over all locations and all weeks) for a fixed classification scheme, *i.e.*, R = 40, and forecast windows (i.e., *N* = 1, 2, 4, 8 and 12), where the risk indicator is case counts.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

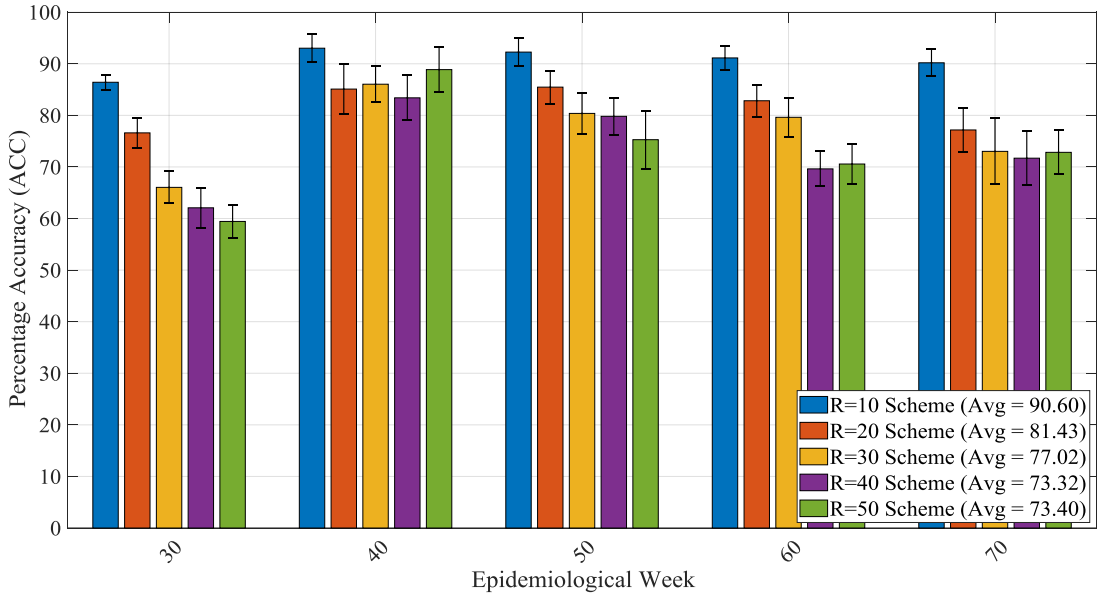
The aggregated ROC curves (averaged over all locations and all epidemiological weeks) are presented in Fig. 6, and reveal the (expected) increased accuracy of the model as the forecast window is reduced. The ROC AUC results are consistent with ACC results presented in Fig. 5, highlighting the superior performance of the 1 and 2 week ahead prediction capability of the model. The ROC AUC value remains above 0.91 for N=1,2 and above 0.83 for N=4, both indicating high predictive accuracy of the model.

Fig. 7 illustrates how the model performance varies throughout the course of the outbreak, presented here for selected epidemiological weeks (i.e., week number / starting date: 30 / 18-Jan-2016, 40 / 28-Mar-2016, 50 / 6-Jun-2016, 60 / 15-Aug-2016, and 70 / 24-Oct-2016). This time period represents a highly complex period of the outbreak with country level rankings fluctuating over time, as evidenced in Fig 1. Fig. 7.a and 7.b present the model performance when different risk indicators are used to classify the countries into high and low risk groups, namely reported case counts and incidence rate, respectively. The mean ACC is reported for a fixed 4-week prediction window, and each by classification scheme. The expected ACC value is averaged over all countries, and the error bars indicate the variability in expected ACC across model runs. The short error bars indicate, critically, the robustness of the model predictions. The model is also demonstrated to perform consistently throughout the course of the epidemic, with the exception of week 30, at which time there was limited information available to train the model, *e.g.*, the outbreak was not yet reported in a majority of the affected countries. Comparing Fig 7.a and 7.b reveals relatively similar performance for both risk indicators, demonstrating the model's flexibility and adaptability with respect to the metric used to classify outbreak risk, *i.e.*, number of cases or incidence rate in a region. Additionally, for both risk indicators, the model accuracy is highest for the more precise classification schemes ($R < 20$), which is consistent with the aggregate model performance illustrated in Fig 5.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a)



(b)

Fig 7. Model performance and variability for selected epidemiological weeks when risk indicator (a) case counts and (b) incidence rate. ACC is averaged over all locations. Combinations of Epi week and classification schemes (i.e., R = 10, 20, 30, 40 and 50), with a fixed forecast windows (i.e., N = 4) are shown. The error bars represent the variability in expected ACC across runs**.**

We further explore the model performance at a regional level by dividing the countries and territories in the Americas into three groups, namely *Caribbean*, *South America* and *Central America*, as in [10]. For each group the average performance of the model in terms of ACC was evaluated and compared. The results in Fig 8 reveals a similar trend at the regional level as was seen at the global level, with a decrease in predictive accuracy as the forecast window increases in length, and the and high risk group increases in size. The results reveal the predictive accuracy is best for the *Caribbean* region, while predictions for *Central America* were consistently the worst; the discrepancy in performance between these groups increases as the forecast

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

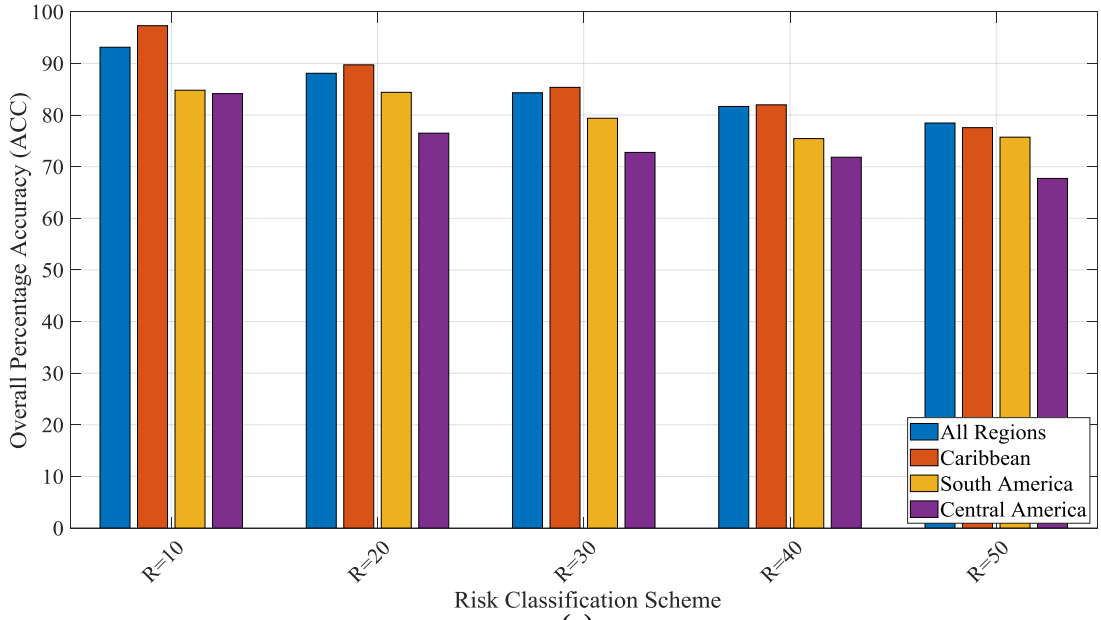
window increases. The difference in performance across regions can be attributed to the high spatial heterogeneity of the outbreak patterns, the relative ability of air travel to accurately capture connectivity between locations, and errors in case reporting that may vary by region. For example, the *Caribbean*, which consists of more than twice as many locations as any other group, first reported cases around week 25, and remained affected throughout the epidemic. In contrast, *Central America* experienced a slow start to the outbreak (at least according to case reports) with two exceptions, namely Honduras and El Salvador. The large number of affected region in the Caribbean, with more reported cases distributed over a longer time period contributed to the training of the model, thus improving the predictive capability for these regions. Additionally, the geographically isolated nature of Caribbean islands enables air travel to more accurately capture incoming travel risk, unlike countries in Central and South America, where individuals can also move about using alternative modes, which are not accounted for in this study. These factors combined explain the higher predictive accuracy of the model for the Caribbean region, and importantly, helps to identify the critical features and types of settings under which this model is expected to perform best.

Conclusions

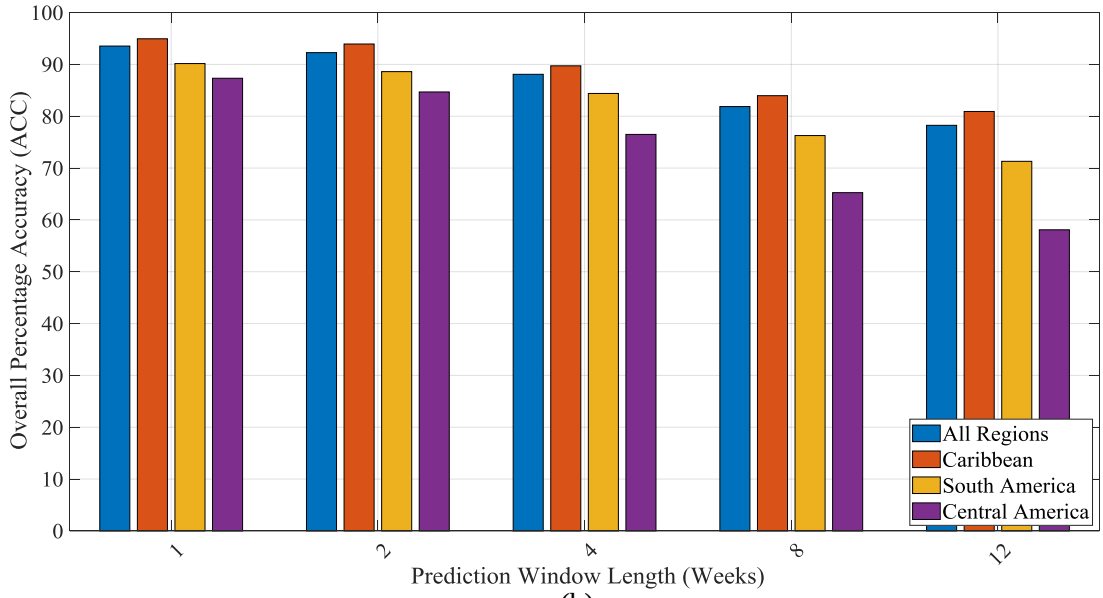
We have introduced a flexible, predictive modelling framework to forecast outbreak risk in real-time. An application of the model was applied to the Zika epidemic in the Americas at a weekly temporal resolution, and country-level spatial resolution, using population, socioeconomic, epidemiological, travel patterns and vector suitability data. The model performance was evaluated for various risk classification schemes, forecast windows and risk indicators, and illustrated to be accurate and robust across a broad range of these features. First, the model is more accurate for shorter prediction windows and restrictive risk classification schemes. Secondly, regional analysis reveals superior predictive accuracy for the Caribbean, suggesting the model to be best suited to geographically isolated locations that are predominantly connected via air travel. Predicting the spread to areas that are relatively isolated has previously been shown to be difficult due to the stochastic nature of infectious disease spread [77]. Thirdly, the model performed consistently well at various stages throughout the course of the outbreak, indicating it's potential value at the early stages of an epidemic. The outcomes from the model can be used to better guide outbreak resource allocation decisions, and can be easily adapted to model other vector-borne epidemics.

There are several limitations of this work. The underlying data on case reporting vary by country and may not represent the true transmission patterns [78]. However, the framework presented was flexible enough to account for these biases and we anticipate will only be improved as data become more robust. Additionally, 2015 travel data was used in place of 2016 data, as has been done previously [43, 58, 59], which may not be fully representative of travel behaviour. Lastly, due to the lack of spatial resolution of case reports, we were limited to make country to country spread estimates. We do however appreciate that there is considerable spatial variation within countries (i.e., northern vs. southern Brazil) and that this may influence the weekly covariates used in this study. We again hypothesise that models will become better as spatial resolution increases.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



(a)



(b)

Fig 8. Regional prediction accuracy for varying (a) risk classification schemes and (b) forecast windows. In (a) the forecast window is fixed to N=4, and in (b) the classification scheme is fixed to R=20. ACC shown for each classification scheme is averaged for the subset of countries in each region over all weeks. The risk indicator used is case counts.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Data Accessibility. All data used in this study is provided as supplementary material.

Competing Interests. We have no competing interests.

Authors' Contributions. LG and MA conceived the study, designed the experiments, analyzed the model results, and drafted the original manuscript. MA developed the model and performed the computational analysis. All authors contributed to data curation and editing of the manuscript. LG supervised the study.

Acknowledgements. We thank Raja Jurdak and Dean Paini for their inputs and discussion on the model.

Funding. We receieved no funding for this work**.**

References

1. Chouin-Carneiro, T., et al., *Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus.* PLoS Negl Trop Dis, 2016. **10**(3): p. 1-11.
2. Dick, G.W., *Zika virus. II. Pathogenicity and physical properties.* Trans R Soc Trop Med Hyg, 1952. **46**(5): p. 521-34.
3. Duffy, M.R., et al., *Zika virus outbreak on Yap Island, Federated States of Micronesia.* N Engl J Med, 2009. **360**(24): p. 2536-43.
4. Hancock, W.T., M. Marfel, and M. Bel, *Zika virus, French Polynesia, South Pacific, 2013.* Emerg Infect Dis, 2014. **20**(11): p. 1960.
5. Dupont-Rouzeyrol, M., et al., *Co-infection with Zika and dengue viruses in 2 patients, New Caledonia, 2014.* Emerg Infect Dis, 2015. **21**(2): p. 381-2.
6. Musso, D., E.J. Nilles, and V.M. Cao-Lormeau, *Rapid spread of emerging Zika virus in the Pacific area.* Clin Microbiol Infect, 2014. **20**(10): p. O595-6.
7. Tognarelli, J., et al., *A report on the outbreak of Zika virus on Easter Island, South Pacific, 2014.* Arch Virol, 2016. **161**(3): p. 665-8.
8. Faria, N.R., et al., *Zika virus in the Americas: Early epidemiological and genetic findings.* Science, 2016.
9. Campos, G.S., A.C. Bandeira, and S.I. Sardi, *Zika Virus Outbreak, Bahia, Brazil.* Emerg Infect Dis, 2015. **21**(10): p. 1885-6.
10. PAHO, *Regional Zika Epidemiological Update (Americas)*, P.A.H.O. World Health Organization, Editor. 2017: Washington DC.
11. Zanluca, C., et al., *First report of autochthonous transmission of Zika virus in Brazil.* Mem Inst Oswaldo Cruz, 2015. **110**(4): p. 569-72.
12. Scott, T.W. and A.C. Morrison, *Vector dynamics and transmission of dengue virus: implications for dengue surveillance and prevention strategies: vector dynamics and dengue prevention.* Curr Top Microbiol Immunol, 2010. **338**: p. 115-28.
13. Achee, N.L., et al., *A critical assessment of vector control for dengue prevention.* PLoS Negl Trop Dis, 2015. **9**(5): p. e0003655.
14. *Vector control with a focus on Aedes aegypti and Aedes albopictus mosquitoes: literature review and analysis of information*. 2017, European Centre for Disease Prevention and Control: Stockholm: ECDC.
15. McGough, S.F., et al., *Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social*

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Media, and News Report Data. PLoS Negl Trop Dis, 2017. **11**(1): p. e0005295.

16. Martínez-Bello, D.A., A. López-Quílez, and A. Torres-Prieto, *Bayesian dynamic modeling of time series of dengue disease case counts.* PLOS Neglected Tropical Diseases, 2017. **11**(7): p. e0005696.
17. Guo, P., et al., *Developing a dengue forecast model using machine learning: A case study in China.* PLOS Neglected Tropical Diseases, 2017. **11**(10): p. e0005973.
18. Johansson, M.A., et al., *Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico.* Sci Rep, 2016. **6**: p. 33707.
19. Earnest, A., et al., *Comparing Statistical Models to Predict Dengue Fever Notifications.* Computational and Mathematical Methods in Medicine, 2012. **2012**: p. 6.
20. Hii, Y.L., et al., *Forecast of Dengue Incidence Using Temperature and Rainfall.* PLOS Neglected Tropical Diseases, 2012. **6**(11): p. e1908.
21. Shi, Y., et al., *Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore.* Environ Health Perspect, 2016. **124**(9): p. 1369-75.
22. Cortes, F., et al., *Time series analysis of dengue surveillance data in two Brazilian cities.* Acta Trop, 2018. **182**: p. 190-197.
23. Abdur Rehman, N., et al., *Fine-grained dengue forecasting using telephone triage services.* Science Advances, 2016. **2**.
24. Lowe, R., et al., *Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador.* Lancet Planet Health, 2017. **1**(4): p. e142-e151.
25. Ramadona, A.L., et al., *Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data.* PLoS One, 2016. **11**(3): p. e0152688.
26. Lauer, S.A., et al., *Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010-2014.* Proc Natl Acad Sci U S A, 2018. **115**(10): p. E2175-E2182.
27. Baquero, O.S., L.M.R. Santana, and F. Chiaravalloti-Neto, *Dengue forecasting in Sao Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models.* PLoS One, 2018. **13**(4): p. e0195065.
28. Sirisena, P., et al., *Effect of Climatic Factors and Population Density on the Distribution of Dengue in Sri Lanka: A GIS Based Evaluation for Prediction of Outbreaks.* PLoS One, 2017. **12**(1): p. e0166806.
29. Anggraeni, W. and L. Aristiani. *Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia*. in *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. 2016.
30. Marques-Toledo, C.A., et al., *Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level.* PLoS Negl Trop Dis, 2017. **11**(7): p. e0005729.
31. Cheong, Y.L., P.J. Leitão, and T. Lakes, *Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees.* Spatial and Spatio-temporal Epidemiology, 2014. **10**: p. 75-84.
32. Wesolowski, A., et al., *Impact of human mobility on the emergence of dengue epidemics in Pakistan.* Proc Natl Acad Sci U S A, 2015. **112**(38): p. 11887-92.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

33. Zhu, G., et al., *Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China.* PLoS Negl Trop Dis, 2016. **10**(4): p. e0004633.
34. Zhu, G., et al., *The spatiotemporal transmission of dengue and its driving mechanism: A case study on the 2014 dengue outbreak in Guangdong, China.* Sci Total Environ, 2018. **622-623**: p. 252-259.
35. Liu, K., et al., *Dynamic spatiotemporal analysis of indigenous dengue fever at street-level in Guangzhou city, China.* PLOS Neglected Tropical Diseases, 2018. **12**(3): p. e0006318.
36. Li, Q., et al., *Spatiotemporal responses of dengue fever transmission to the road network in an urban area.* Acta Trop, 2018. **183**: p. 8-13.
37. Chen, Y., et al., *Neighbourhood level real-time forecasting of dengue cases in tropical urban Singapore.* BMC Medicine, 2018. **16**(1): p. 129.
38. Gardner, L. and S. Sarkar, *A global airport-based risk model for the spread of dengue infection via the air transport network.* PLoS One, 2013. **8**(8): p. e72129.
39. Lauren M. Gardner, D.F., S. Travis Waller, Ophelia Wang and Sahotra Sarkar, *A Predictive Spatial Model to Quantify the Risk of Air-Travel-Associated Dengue Importation into the United States and Europe.* Journal of Tropical Medicine, 2012. **2012**.
40. Grubaugh, N.D., et al., *Genomic epidemiology reveals multiple introductions of Zika virus into the United States.* Nature, 2017. **546**: p. 401.
41. Wilder-Smith, A. and D.J. Gubler, *Geographic expansion of dengue: the impact of international travel.* Med Clin North Am, 2008. **92**(6): p. 1377-90, x.
42. Faria, N.R., et al., *Zika virus in the Americas: Early epidemiological and genetic findings.* Science, 2016. **352**(6283): p. 345-349.
43. Gardner, L.M., et al., *Inferring the risk factors behind the geographical spread and transmission of Zika in the Americas.* PLoS Negl Trop Dis, 2018. **12**(1): p. e0006194.
44. Tatem, A.J. and S.I. Hay, *Climatic similarity and biological exchange in the worldwide airline transportation network.* Proceedings of the Royal Society B: Biological Sciences, 2007. **274**(1617): p. 1489.
45. Siriyasatien, P., et al., *Analysis of significant factors for dengue fever incidence prediction.* BMC Bioinformatics, 2016. **17**(1): p. 166.
46. Nishanthi P h m Herath, A.a.i.P.a.H.p.W., *Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks.* International Journal of Computer Applications, 2014. **101**(15): p. 1-5.
47. Aburas, H.M., B.G. Cetiner, and M. Sari, *Dengue confirmed-cases prediction: A neural network model.* Expert Systems with Applications, 2010. **37**(6): p. 4256-4260.
48. Baquero, O.S., L.M.R. Santana, and F. Chiaravalloti-Neto, *Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models.* PLOS ONE, 2018. **13**(4): p. e0195065.
49. Faisal, T., M.N. Taib, and F. Ibrahim, *Neural network diagnostic system for dengue patients risk classification.* J Med Syst, 2012. **36**(2): p. 661-76.
50. Laureano-Rosario, E.A., et al., *Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan,*

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mexico and San Juan, Puerto Rico. Tropical Medicine and Infectious Disease, 2018. **3**(1).

51. Kiskin I, O.B., Windebank T, Zilli D, Sinka M, Willis K, Roberts S, *Mosquito detection with neural networks: the buzz of deep learning.* arXiv, 2017.
52. Scavuzzo, J.M., et al. *Modeling the temporal pattern of Dengue, Chicungunya and Zika vector using satellite data and neural networks*. in *2017 XVII Workshop on Information Processing and Control (RPIC)*. 2017.
53. Sanchez-Ortiz, A., et al. *Mosquito larva classification method based on convolutional neural networks*. in *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*. 2017.
54. Nguyen, T., et al. *Epidemiological dynamics modeling by fusion of soft computing techniques*. in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. 2013.
55. Jiang, D., et al., *Mapping the transmission risk of Zika virus using machine learning models.* Acta Tropica, 2018. **185**: p. 391-399.
56. Wahba, G., *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. 1990: Society for Industrial and Applied Mathematics. 177.
57. PAHO. *Countries and territories with autochthonous transmission in the Americas reported in 2015-2017*. 2017; Available from: http://www.paho.org/hq/index.php?option=com_content&view=article&id=11603&Itemid=41696&lang=en.
58. Gardner, L., N. Chen, and S. Sarkar, *Vector status of Aedes species determines geographical risk of autochthonous Zika virus establishment.* PLoS Negl Trop Dis, 2017. **11**(3): p. e0005487.
59. Gardner, L.M., N. Chen, and S. Sarkar, *Global risk of Zika virus depends critically on vector status of Aedes albopictus.* Lancet Infect Dis, 2016. **16**(5): p. 522-523.
60. Kraemer, M.U., et al., *The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus.* Elife, 2015. **4**: p. e08347.
61. WorldBank. *International Comparison Program database. GDP per capita, PPP*. 2016; Available from: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD.
62. Analysis, U.S.B.o.E. *Widespread Economic Growth Across States In 2011*. 2011.
63. Services, U.S.D.o.H.a.H. *Health, United States, 2015* 2015; Available from: https://www.cdc.gov/nchs/data/hus/hus15.pdf.
64. (WHO), W.H.O. *WHO World Health Statistics 2015*. 2015; Available from: http://www.who.int/gho/publications/world_health_statistics/2015/en/.
65. PAHO. *PLISA Health Indication Platform for the Americas*. 2017; Available from: http://www.paho.org/data/index.php/en/.
66. Bank, W. *International Comparison Program database. Population density (people per sq. km of land area)*. 2016; Available from: http://data.worldbank.org/indicator/EN.POP.DNST.
67. *International Air Travel Association (IATA)- Passenger Intelligence Services (PaxIS).* Available from: http://www.iata.org/services/statistics/intelligence/paxis/Pages/index.aspx.
68. Pigott, D., et al., *Local, national, and regional viral haemorrhagic fever pandemic potential in Africa: a multistage analysis.* Lancet, 2017. **390**(10113): p. 2662-2672.

bioRxiv preprint doi: https://doi.org/10.1101/466581; this version posted November 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

69. Leontaritis, I.J. and S.A. Billings, *Input-output parametric models for non-linear systems Part I: deterministic non-linear systems.* International Journal of Control, 1985. **41**(2): p. 303-328.
70. Narendra, K.S. and K. Parthasarathy, *Identification and control of dynamical systems using neural networks.* IEEE Transactions on Neural Networks, 1990. **1**(1): p. 4-27.
71. Chen, S., S.A. Billings, and P.M. Grant, *Non-linear system identification using neural networks.* International Journal of Control, 1990. **51**(6): p. 1191-1214.
72. Siegelmann, H.T., B.G. Horne, and C.L. Giles, *Computational capabilities of recurrent NARX neural networks.* IEEE Trans Syst Man Cybern B Cybern, 1997. **27**(2): p. 208-15.
73. Tsungnan, L., et al., *Learning long-term dependencies is not as difficult with NARX recurrent neural networks*. 1995, University of Maryland at College Park. p. 23.
74. *MATLAB and Neural Network Toolbox Release 2018a*. [cited 2018 16 July]; Available from: https://au.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf.
75. Boussaada, Z., et al., *A Nonlinear Autoregressive Exogenous (NARX) Neural Network Model for the Prediction of the Daily Direct Solar Radiation.* Energies, 2018. **11**(3).
76. Fawcett, T., *ROC graphs: Notes and practical considerations for researchers.* Machine Learning, 2004. **31**: p. 1-38.
77. Brockmann, D. and D. Helbing, *The Hidden Geometry of Complex, Network-Driven Contagion Phenomena.* Science, 2013. **342**: p. 1337-1342.
78. Faria, N.R., et al., *Establishment and cryptic transmission of Zika virus in Brazil and the Americas.* Nature, 2017. **546**: p. 406.