Journal of the Indian Statistical Association
Vol.?? No. ? & ?, 20??,

Spatio-Temporal Biosurveillance of Climate Sensitive Mosquito-Borne Diseases Using Online Social Media

L. Leticia Ramirez Ramirez

Centro de Investigacion en Matematicas, Mexico

Vyacheslav Lyubchich

University of Maryland Center for Environmental Science, USA

Yulia R. Gel

University of Texas at Dallas, USA

Abstract

Chikungunya, as many other climate-sensitive, vector-borne infectious diseases, has recently re-emerged in the Americas. Since its first detection in 1952 in Tanzania, chikungunya virus has spread over 60 countries in Europe, Africa, Asia, and the Indian subcontinent, but the first outbreak in the Americas was reported only in late 2013. One of the primary challenges for modeling and predicting emerging vector-borne infectious diseases is short or even non-existent historical data records. Even if available, the information is often highly noisy, incomplete, or reported with delays, which obstructs any prediction task. We propose a new integrative multi-source and computationally efficient statistical approach for out-of-sample prediction of emerging vector-borne infections (with an emphasis on predicting the epidemic peak and duration) that is applicable even when epidemiological data are limited. The strength of the new method is based on the two pillars at the interface of statistics, applied mathematics and data science. First, we systematically integrate multi-source information, from official public health records to non-traditional bio surveillance online social media data. Second,

Received : month year

we propose a comprehensive modeling framework, spanning ordinary differential systems to model the infection spread between humans and mosquitoes, and statistical approaches such as time series clustering and the Box–Jenkins forecasting methodology. We illustrate our approach in application to 5-month ahead forecasting of chikungunya in the Dominican Republic.

Key Words : *biosurveillance, chikungunya, dengue, environmental risk, social media, spatio-temporal clustering*

1 Introduction

Chikungunya is a disease caused by an *alphavirus* (CHIKV) that is mainly transmitted by two mosquito species, namely, *Aedes aegypti* and *Aedes albopictus*. The same species transmit other mosquito-borne viruses such as dengue, yellow fever, and Zika virus [28, 4, 27]. Currently there exists neither specific antiviral drug nor chikungunya vaccine, and its medical treatments primarily aim to help relieve the symptoms.

Chikungunya was first detected during an outbreak in 1952 in Tanzania and later it was reported in Europe, Southeast Asia, India, and islands in the Indian and Pacific Oceans, with a major outbreak in Réunion Island in 2006. Since 2004, chikungunya expanded its geographical presence, causing sustained epidemics of unprecedented magnitude in Asia and Africa. The virus outbreaks were observed also on the islands in the Indian Ocean and in Italy, although these areas are considered to be endemic for this disease.

The number of laboratory-confirmed or probable cases of chikungunya in the Americas increased from a few, in 1995–2005, to 106 cases in 2006–2010. However, all these identified cases were imported and none of them led to local transmission [4].

In December 2013, two laboratory-confirmed cases of chikungunya without a travel history were reported on the French part of the Caribbean island of Saint Martin, indicating the start of the first documented outbreak

of chikungunya in the Americas [37]. Afterwards, the agent spread rapidly over Caribbean and into North, Central, and South America. Over the period of 2013–2014, there were 1,071,696 reported suspected cases and 22,796 confirmed cases of chikungunya in the Americas [27].

Emerging diseases in a region is a cause of local and international concern, especially given a strong focus of the national economies on tourism. The 2013 chikungunya outbreak attracted the attention from the healthcare professionals in North, Central, and South Americas. Moreover, on 15 August 2014 the Defense Advanced Research Projects Agency (DARPA) of the USA launched a chikungunya forecasting challenge, with a goal to develop new data analytics tools for predicting chikungunya dynamics in the Americas [11].

Under the extreme scenarios of emerging diseases, as the case of chikungunya, official data capture only few recent weeks of disease activity in a region. This paper addresses this challenge by proposing an innovative comprehensive integration of information from different sources in a statistical model, with the objective of obtaining a long-term forecast to assess the most crucial epidemic period (that is, epidemic peak) and the outbreak duration. We propose obtaining several preliminary forecasts, each derived from a different information source, to be integrated into a statistical model that provides a strengthened forecast.

Regarding the alternative data on chikungunya, we note that this infectious agent is transmitted by the same species of mosquito as dengue. Hence, we propose to use online social media data on dengue, specifically, Google Dengue Trends (GDT), as proxy information to the unreported number of cases, the unknown mosquito density, and interaction with humans. In particular, the lack of reliable timely available public health records is typically one of the primary challenges obstructing real-time epidemic forecasting, and this problem is particularly acute in developing countries. The key contribution of such social media data is to deliver

epidemiological signal in the absence of the confirmed traditional offline information. In our study we focus on the chikungunya outbreak in the Dominican Republic (DR), as DR offers one of the most consistent chikungunya records. However, GDT data are unavailable for this country. Therefore, as a preliminary step before obtaining a forecast based on GDT, we apply data mining techniques to identify areas in Mexico – the country with the highest spatial resolution of GDT in Latin America – that exhibit the highest level of correspondence with the chikungunya records for the Dominican Republic.

Non-traditional biosurveillance data from online social media sources such as Google, Twitter, Facebook, and Wikipedia, are criticized for high sensitivity to self-excitement (i.e., fickle media interest), bias and other artifacts of social media [21]. To address this challenge, we combine the GDT information with predictions based on an epidemic compartmental model fitted with the available official public reports as well as adaptively re-calibrate social media data with the offline data. The historical GDT for the selected states in Mexico and the estimated curves from the compartmental model are introduced as exogenous covariates into a time series model for predicting the chikungunya epidemiological curve (i.e., location and intensity of spike).

The idea of combining different information sources for improving infectious disease modeling has been studied by, for example, [10] and [31]. However, only [24] proposed a multi-source forecast for a vector-borne infectious disease (Zika) for up to one, two, and three weeks ahead.

The earliest epidemic models for chikungunya forecast in the Americas are associated to the DARPA chikungunya challenge [11] that asked participants to forecast the cumulative total cases (suspected and confirmed, the latter including imported-confirmed) per week per country in the Americas. Participants typically complemented the information provided by Pan American Health Organization (PAHO) with data from other

sources, including online web searches, climate information, and vector-specific information (e.g., reporting of other mosquito-borne illnesses such as dengue in the same population, mosquito dynamics, and ecology). Forecasts with higher accuracy used between 1 and 8 data sources, but overall there was no significant correlation between the number of data sources and the accuracy of the forecast. In fact, not all data sources were considered in deriving the final prediction, and the four top ranking forecasts where based on the morphological approach.

[13] proposed a model that uses the locally originated and imported cases, based on the ecology for the mosquitoes niche, and air travel routes. This approach incorporates information on atmospheric variables and airlines routes, but it still highly depends on the official surveillance data accuracy and timeliness. [13, 17] modeled chikungunya spread by modeling the local and imported cases. Based on two branching processes, for the imported cases and local transmissions, and information on travel origin-destinations and averaged environmental variables, the authors provide probabilistic nowcasting for chikungunya in the Americas.

In contrast to the top ranking forecast for the DARPA challenge, we propose a model that is not morphological, but is a combination of traditional epidemic and statistical models. Opposing to [13] and [17], we base the forecast for CHIKV in DR on the local transmission and we do not consider environmental variables to model the mosquito density, but we approximate the spread of mosquitoes and the mosquito-human interactions using historical GDT for some states in Mexico. As [13], but contrary to [17] (and [24], for Zika), we aim to obtain long-term forecast (five months) at early stages of an emerging epidemic outbreak. Finally, in contrast to [13], we are able to obtain accurate predictions for location and intensity of spike using only the first eight weeks into the outbreak.

The remainder of the paper describes the used information (Section 2), then introduces the general methodology for the integrated model

(Section 3). Sections 4 presents the results of forecasting the 2014 chikungunya outbreak in DR, and Section 5 contains the discussion.

2 Data description and challenges

As a result of the early outbreak warning issued by PAHO in December 2013, countries in the Americas, as Dominican Republic, reinforced their surveillance system in 2014. The collected information was published by PAHO in electronic weekly bulletins available online (for Dominican Republic, see [27]). The PAHO bulletins report the cumulative numbers of suspected and confirmed cases on a weekly basis. While the reports undergo standard PAHO procedures on data quality control, there are numerous delays in reporting new chikungunya cases; furthermore, occasionally information on multiple weeks is aggregated, leading to the so-called plateaus. In such instances, to preserve weekly time granularity in reporting, we retrospectively uniformly split the reported data among weeks for which information has not been updated. Since the population was not previously exposed to this virus, the objective of our model is predicting the epidemic curve for the total number of reported suspected and confirmed cases.

Since the official information on emerging diseases can be very scarce, we complement the official reports with previous activity of Internet-based disease monitoring. In particular, we use Google Dengue Trends (GDT). These sources of information have been studied in applications on emerging or vector-borne diseases. [5] proposed using web search queries to estimate the dengue prevalence in endemic countries, where the surveillance systems fail to timely report the suspected and confirmed cases. [25] reviewed studies that had exploited Internet use and search trends to monitor influenza and dengue. The authors conclude that the web searches (Google Influenza and Dengue Trends) show a potential in reinforcing traditional surveillance system capacity and guiding public health action. Using GDT, [39] proposed

a methodological framework to obtain close to real-time estimates for dengue in Mexico, Brazil, Thailand, Singapore, and Taiwan. These estimates seek to improve the tracking of dengue activity. According to [12] and [14], the dengue season varies from year to year but tends to coincide with the rainy season. For the countries with favorable climate for the vector, GDT is accurate. This observation is also shared by [34] for dengue in Venezuela. In relation to chikungunya, [3] explored Google Trends (GT) correlation with chikungunya and Zika during 2014 in Venezuela and concluded that GT can be used to forecast some outbreak characteristics such as the relative magnitude and duration.

3 Proposed methodology

Our methodology of predicting chikungunya dynamics is based on 1) robust proxy time series, such as aggregates of the GDT in Mexican states, and 2) knowledge of the general epidemics dynamics captured with fitted solutions of an epidemic model given by a system of ordinary differential equations (ODEs). Both GDT proxies and ODEs solutions are then combined in a single time series model with a developed apparatus for making predictions with a desired confidence level (Figure 3).

3.1 Spatio-temporal clustering

Whereas the dengue and chikungunya viruses are known to be mainly transmitted by the same mosquito species, and long GDT records are available for Mexico, it would be incorrect to directly use the Mexican GDT data to model chikungunya cases in Dominican Republic. First, correlation between the two countries might not be very strong since they are separated by the sea and do not share a common land border. Second, aggregating data over the vast territory of Mexico (compared with the territory of Dominican Republic) would average out the diverse trends observed in different Mexican

states. Hence, we propose to use GDT data at a finer spatial scale (at the scale of one state or a small group of Mexican states) to select proxies for the chikungunya dynamics in Dominican Republic. To identify those groups of states, we apply spatio-temporal clustering methods.

Due to very diverse topography and climate of Mexico, states with similar traits important for survival of mosquitoes and spread of dengue virus can be spatially separated (i.e., spatial proximity in the clustering becomes less important than similarity of time trends). One of the methods that do not explicitly use the spatial information is the TRend based clUstering algorithm for Spatio-Temporal data stream (TRUST) [7, 15, 1]. TRUST groups spatial data within relatively short temporal intervals (called *slides*) based on a number of homogeneity thresholds. The cluster assignments for each location in each slide can then be used at a higher level of temporal aggregation to cluster the trends within a long temporal *window*. Both slide and window sizes are user-defined.

For forecasting chikungunya, we are interested in selecting GDT series that would be good proxies for the whole epidemic curve of chikungunya, hence, we choose the slides to be of one-year length. To obtain robust combinations of proxy series, we cluster the GDT within the whole period with complete data records (2009–2013), i.e., we set the window size to be five years. The two levels of clustering are schematically shown in Figure 1.

The slide-level clustering starts with identifying the binary δ-close measure for the pairs of time series x_i and x_j $(i, j = 1, \ldots, n;\ i \neq j)$:

$$f_\delta(x_i, x_j) = \mathbb{1}\left(\frac{\|x_i - x_j\|_1}{T_1 - T_0} \leq \delta\right),$$

where $[T_0, T_1]$ is the time series domain and $\delta \in [0, 1]$. At the final step, window-level clustering can be defined with

$$f_\epsilon(x_i, x_j) = \mathbb{1}\left(M^{-1}\sum_{s=1}^{M} \text{clustered}(x_i, x_j, s) \geq \epsilon M\right),$$

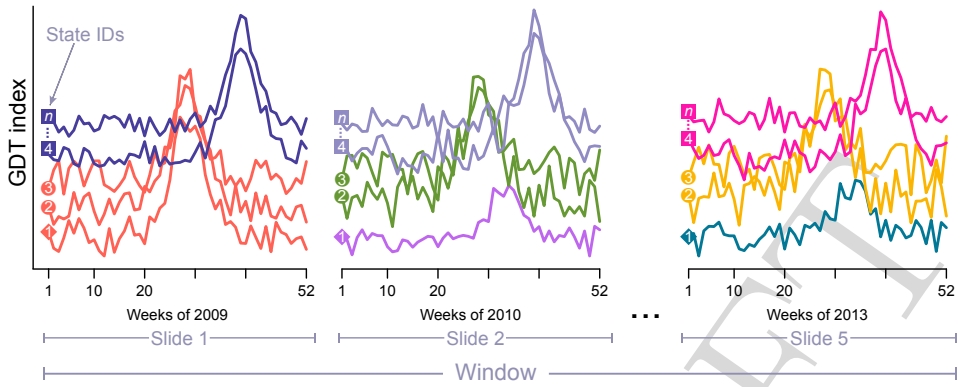

Figure 1: Scheme of the TRUST algorithm for clustering GDT in the states of Mexico in a window of 2009–2013. Colors identify slide-level clusters; shapes around the state IDs denote window-level clusters.

where M is the number of slides in a window, $\epsilon \in [0, 1]$, and clustered(x_i, x_j, s) is a binary variable taking on the value of 1 if the two time series are clustered in the s-th slide, and the value of 0 otherwise (see detailed discussion of the TRUST algorithm in [7]). We select the tuning parameters δ and ϵ in a data-driven way using a grid search with Bayesian information criterion [32, 23].

After the window-level clusters of GDT are obtained, the proxy for chikungunya is identified as a cluster-averaged time series that has the highest correlation with the chikungunya time series.

3.2 Inverse problem of an ODE epidemic model

In addition to the information collected from GDT, we use forecasts derived from a compartmental deterministic epidemic model fitted with the available official surveillance reports. The epidemic model consists of a set of ordinary differential equations parameterized according to $\boldsymbol{\theta}$ where its solution (*forward map*, FM) $G(t; \boldsymbol{\theta})$ describes the evolution, over time, of the infectious agent, in relation to its spread in two interacting populations: humans and mosquitoes.

The compartments of such deterministic epidemic model are defined by the individuals' ability to become infected or transmit the pathogen, such as being susceptible (S), exposed (E), infectious (I), and removed (R). For instance, two relevant compartmental models are SIR and SEIR [18, 19, 20, 2]. We consider a two-population SEIR model, where an individual in the susceptible human population can be exposed to the pathogen, when bitten by an infectious mosquito. At the end of this period the person becomes infectious (being able to pass the virus to mosquitoes), and later recovers, attaining temporal immunity to the agent. The number of humans in each of the stages are denoted as S, E, I and R, respectively. We consider an SEI model for mosquitoes, with the respective stages denoted by X, Y, and Z.

In the Bayesian approach to the inverse problem [8, 9, 35, 30, 36], we assume that the surveillance reports $\boldsymbol{y} = y_1, \ldots, y_m$ are a modification and/or imperfect observation of the forward map. That is,

$$Y_i = h(w_i(G(t; \boldsymbol{\theta}_0)); \boldsymbol{\alpha}), \quad i = 1, \ldots, m,$$

where $\boldsymbol{\theta}_0$ is the "true" epidemic model parameter that we aim to estimate. w_i is a function of all the states of the "real" compartmental model, and h, along with $\boldsymbol{\alpha}$, describes the deviations from the model originated due to the stochastic nature of the infectious dynamic. In the specific case of epidemic models, the function w_i usually ignores the number of cases in other states (E for example) and only reports the number of symptomatic infected individuals. On top of this modification, we also consider that the data is aggregated by time periods (then i corresponds to the index for the i-th week). These modifications seek to capture the most important features of the real epidemic surveillance and reporting systems. On the other hand, there exist multiple reasons why the reporting data can vary from $w_i(G(t; \boldsymbol{\theta}_0))$. Some of these can be the possible different individual times to develop symptoms, times to look for medical attention after disease

onset, etc. To model these random deviations, we use h and its parameter vector $\boldsymbol{\alpha}$, and we propose the random model to have $w_i(G(t;\boldsymbol{\theta}_0))$ as expected value. Then we assume that under $\boldsymbol{\theta}_0$, the reported observation y_i is the realization of the random variable Y_i, for i, any of the m observed weeks.

A very important source of variation is also the sub-reporting, that is the under representation for the true number of cases. This is mainly owing to the fact that some individuals can experience mild symptoms and do not seek for medical attention. The under-reporting can be exacerbated when surveillance is partially implemented due to limited resources.

Since we have a deterministic system subjected to diverse sources of uncertainty, the predictions based on the surveillance reports and the epidemic FM, falls within the realm of uncertainty quantification. The goal is to use the observations $\boldsymbol{y}$ to deduce the value of the parameter $\boldsymbol{\theta}_0$ in the underlying model $G(t;\boldsymbol{\theta}_0)$, and the parameter $\boldsymbol{\alpha}$ that describes the data variation. Under the Bayesian paradigm, the likelihood corresponds to a statistical model that introduces the FM and the deviation from it present in the observed values.

Noting that the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha})$ is a function of the FM, which in the majority of the cases has to be numerically obtained, we turn to Markov chain Monte Carlo (MCMC) methods to obtain samples of the posterior distribution $p(\boldsymbol{\theta},\boldsymbol{\alpha}|\boldsymbol{y})$. Based on this distribution, we can obtain a point estimate $\hat{\boldsymbol{\theta}}$ to draw an epidemic prediction from $G(t;\hat{\boldsymbol{\theta}})$, or use information on the posterior distribution to obtain an estimated predictive distribution. In either case, the FM evaluated in $\hat{\theta}$, or the ensemble of FM's evaluated in the sampled values of $p(\boldsymbol{\theta}|\boldsymbol{y})$, can be combined with the GDT information in a Box–Jenkins time series model.

We propose using the epidemic model for chikungunya virus presented by [38] for the large outbreak in the Réunion Island in 2005–2006. This compartmental two-population SEIR model considers that infectious period starts before the disease and that not all infected humans will develop

noticeable symptoms. While all exposed individuals will become infective, some of them will present the disease or symptoms (I_s), and some will have mild or no symptoms (I_a).

Figure 2 depicts the mathematical model (3.1), where the compartments related to the virus transmission are in white. The compartment D comprises the individuals who develop symptoms. Given the individuals' SEIR status, this compartment is irrelevant for the epidemic dynamic, but it is directly related to the information the surveillance system collects.

$$\begin{aligned} dS &= -\beta_1 SZ \\ dE &= \beta_1 SZ - \lambda_1 E \\ dI_s &= \phi\lambda_1 E - \gamma I_s \\ dD &= \omega I_s \\ dI_a &= (1-\phi)\lambda_1 E - \gamma I_a \\ dR &= \gamma(I + I_a) \\ dX &= \mu - \beta_2 X(I_s + I_a) - \mu X \\ dY &= \beta_2 X(I_s + I_a) - \lambda_2 Y - \mu Y \\ dZ &= \lambda_2 Y - \mu Z \end{aligned} \tag{3.1}$$

The fraction ϕ of exposed individuals who become infectious will develop symptoms at a rate ω after entering I_s. Note that transiting to D does not imply the ending of the infectious period for infectious individuals since the transition to R is dictated by the rate γ, for both symptomatic and asymptomatic individuals in $I_s \cup I_a$.

Considering outbreak evolution occurs within a year or less, we assume that human demographic changes are negligible. This is not the case for mosquitoes, in view of their short life span. Mosquitoes can live as adults only from two weeks to a month, depending on environmental conditions. [38] introduced the mortality rate of mosquitoes as μ at any of their SEI stages (see Figure 2), that is, mosquitoes' deaths are modeled as independent to the infection status. Following [38], the birth rate of

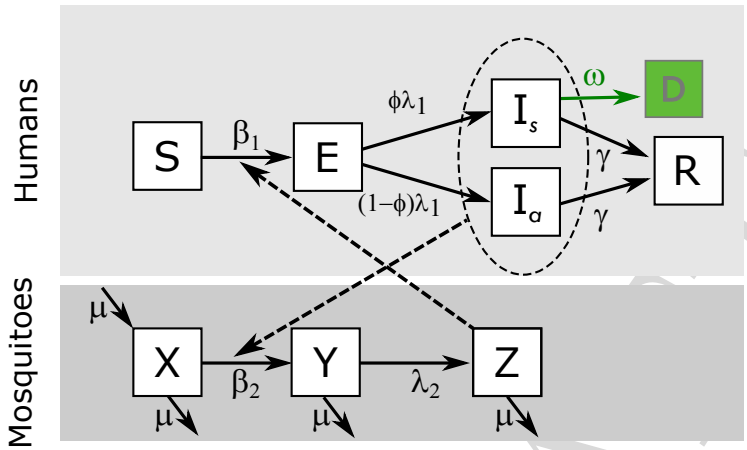

Figure 2: Epidemic Forward Map model.

mosquitoes is also considered equal to μ, which corresponds to a stable mosquito population – something that is observed in tropical environments such as Dominican Republic, where air temperature and humidity are almost constant throughout the year.

The surveillance information is modeled as the weekly number of individuals entering state D, denoted as $x_i = w_i(G(t;\boldsymbol{\theta}))$, for each week i. To describe the deviations of the observed number of new cases y_i from x_i we model Y_i as a random variable. A linear model could be considered, where we add an additive white noise to $\{x_i\}$, however we opt to comply with the discrete nature of the data and propose a model for counts. A natural candidate is a Poisson model with means equal to x_i. Since x_i is the aggregated solution for D in (3.1), it is not necessarily an integer, however a realization of a Poisson with this mean, always is. Nevertheless, this model tends to underestimate real observed variations and the variance cannot be adjusted to describe the real variance in the data. Then we turn to the Negative binomial distribution with mean x_i and variance proportional to

the mean:

$$Y_i \sim \text{NegBin}(x_i/(\alpha-1), 1/\alpha), \tag{3.2}$$

where $\boldsymbol{\theta} = (\beta_1, \lambda_1, \phi, \gamma, \omega, \beta_2, \lambda_2, \mu)$. Then Y_i has mean x_i and variance αx_i.

3.3 The joint modeling framework

The predictive epidemic model is built under the restriction that we observe only the first m weeks of the outbreak, where $y_1, \ldots, y_m$ correspond to the weekly new cases. From the present time m, we aim to obtain $\hat{y}_{m+1}, \ldots, \hat{y}_n$ for some $n > m+1$ and predict some important features in the future epidemic curve, such as the peak position and the outbreak span.

To construct integrated forecasts, we propose to use tractable and computationally efficient multivariate regression with the Box–Jenkins approach of autoregressive integrated moving average modeling (MR-ARIMA). The design matrix X_t comprises the covariates: D_i is the weekly GDT in the selected cluster of Mexican states in the previous year, and $x_i = w_i(G(t;\boldsymbol{\theta}))$ the FM epidemic curve, of the current year, with parameter $\boldsymbol{\theta}$ that can correspond a point estimate or a value sampled the posterior distribution of $\boldsymbol{\theta}$.

To avoid multicollinearity, the proxy $\{D_i\}$ is the averaged series from a cluster that has the highest correlation of this average with $y_1, \ldots, y_m$. The autoregressive order p and moving average order q for the time series of regression errors ϵ_t are selected based on the minimal value of Akaike information criterion (AIC), and the integration order d is selected based on Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (see the algorithm `auto.arima` by [16]). Note that $d > 0$ implies that the following model is fitted not on the original levels of response variable and predictors, but on their respective differences of order d:

$$\begin{aligned} \hat{y}_t &= X_t^\top \boldsymbol{\beta} + \epsilon_t, \\ \epsilon_t &= \sum_{i=1}^{p} \phi_i \epsilon_{t-i} + \sum_{j=1}^{q} \psi_j v_{t-j} + v_t, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, $v_t \sim \text{WN}(0, \sigma^2)$; $p, d, q \in \mathbf{Z}^+$; and ϕ_i and ξ_i are chosen such that $\phi(\lambda) = 1 - \phi_1\lambda - \ldots - \phi_p\lambda^p$ and $\psi(\lambda) = 1 + \psi_1\lambda + \ldots + \psi_q\lambda^q$. Finally, we assume that polynomials $\phi(\cdot)$ and $\psi(\cdot)$ are such that $\phi(\lambda), \psi(\lambda) \neq 0$, $\forall|\lambda| \leqslant 1$, that is, the standard assumptions of stationarity and invertibility conditions in time series.

Then using the information of the first m weeks we obtain the integrated forecast $\hat{y}_i$ for $i = m+1, \ldots, n$ using the design matrix $X_t^\top = (D_i, x_i)^\top$. Figure 3 summarizes this process and the notation we have introduced.

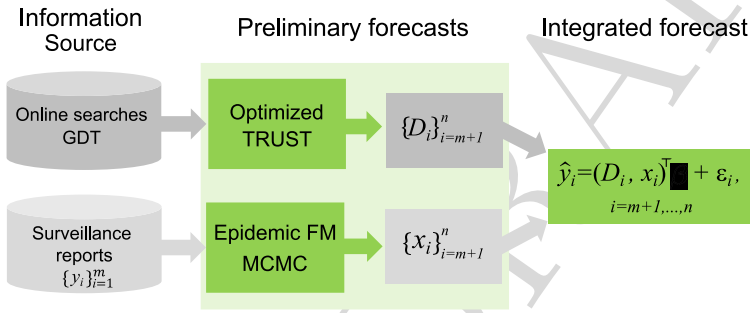

Figure 3: Flow-chart of obtaining the integrated forecast. GDT data (Mexico) and the reported chikungunya cases $\{y_i\}_1^m$ (Dominican Republic) are used to produce preliminary forecasts D_i and x_i ($i = m+1, \ldots, n$), which are then integrated in the MR-ARIMA model to produce provides the final forecast $\{\hat{y}_i\}_{m+1}^n$.

4 Case study: Predictive Analytics for Chikungunya in the Dominican Republic

4.1 Clustering results

The optimized TRUST procedure identified two clusters of states that have consistently exhibited similar dynamics of GDT – Cluster 1: Baja California and Jalisco, and Cluster 2: Tamaulipas, Nayarit, and Tabasco (Figure 4). The other 12 states showed individually different dynamics and did not form clusters. The BIC-selected optimal clustering tuning parameters are δ of

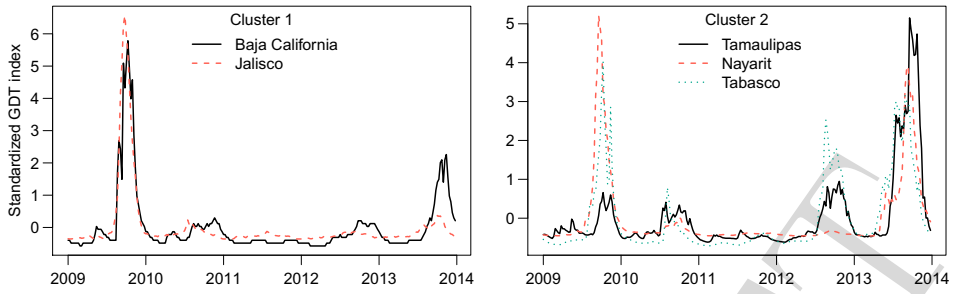

Figure 4: Time series clusters of GDT in Mexico during 2009–2013.

0.167 and ϵ of 1.

To identify which cluster is a reliable proxy for predicting the chikungunya outbreak, we use correlations between chikungunya in the eight non-zero weeks of y_i within $1, 2, \ldots, 23$ weeks of 2014 ($m = 23$), and cluster-average GDTs in the corresponding weeks. We proceed with the cluster exhibiting the highest correlation (Cluster 2, $r = 0.94$, with p-value < 0.01), and continue using its weekly averaged 2013 GDT in the forecasting model (3.3) (from week 16 to 43, Figure 5). Cluster 1 showed non-significant correlation with the chikungunya time series, with correlation r of 0.50 and p-value of 0.21.

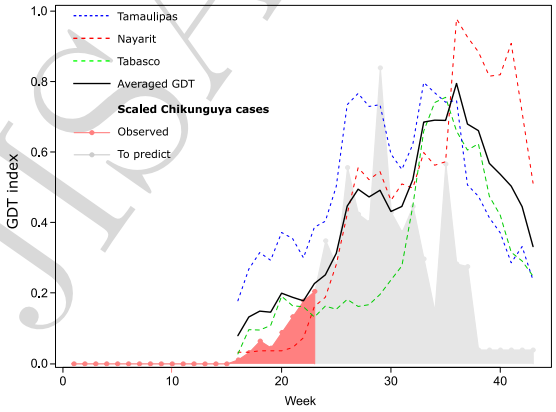

Figure 5: GDT for Cluster 2 for weeks 16–43, 2013.

4.2 Prediction with an epidemic deterministic model subjected to uncertainty

To construct the prior distribution $p(\boldsymbol{\theta})$ we assume that our knowledge on the parameters are independent and modeled with uniform distribution. The distributions of parameters β_1, λ_1, ω, γ, μ, β_2, and λ_2 are centered at 0.14, 0.5, 0.25, 0.25, 0.05, 0.40 and 0.5, respectively. These values correspond to the least squares estimates used in [38] and to information on ranges of parameters in clinical and entomological literature.

Since we do not want to have very informative priors, in contrast to [38], who run sensitivity analysis varying the parameters by 10% their point estimates, we set the domains for our prior as ±30% their midpoints.

In the case of the parameter ϕ, we establish a distribution that also takes into consideration the under-reporting in the surveillance system. That is, we want ϕ to represent the percentage of individuals who develop symptoms and are captured by the surveillance system. Under-reporting is a common problem for countries where the systems of reporting and diagnosis are still to be consolidated. For the Dominican Republic, we assume that the reporting of symptomatic cases can be between 5% and 100%. This interval includes the estimated reporting by [26] for the 2014 chikungunya outbreak in Colombia (39%) and originates the (almost) non-informative prior distribution $\phi \sim$ Unif(0.034,1).

In relation to α, we usually have over-dispersed surveillance count observations, so we establish a flat prior distribution for α. This is modeled as uniform distribution on (1, 800).

Based on the first 23 weeks of 2014 (on which we only observe 8 in the outbreak) and using the MCMC algorithm *twalk* [6] implemented in R [29], we obtain a sample of 370 points from the posterior distribution, after obtaining a chain of 200,000 simulations, examining the trace plots and selecting a conservative burn in period of 170,000. The summary statistics (Table 1) show that the posterior distribution of ϕ concentrates

in values much smaller than 0.97, indicating a strong effect of under-reporting. That is, we estimate $\boldsymbol{\theta}$ and α in relation to the true but under-reported symptomatic cases, to predict the number of cases captured by the surveillance system.

Table 1: Summary of sampled posterior distribution of the ODE parameters

	β_1	λ_1	ϕ	ω	γ	μ	β_2	λ_2	α
2.5%	0.1023	0.3568	0.0326	0.1767	0.1772	0.0358	0.3178	0.3635	441.03
50%	0.1406	0.4898	0.0454	0.2228	0.2423	0.0481	0.4125	0.5198	562.66
97.5%	0.1769	0.6412	0.2174	0.3198	0.3209	0.0639	0.5023	0.6426	598.59
Mean	0.1403	0.4959	0.0647	0.2316	0.2484	0.0490	0.4117	0.5124	554.67

For each of the sampled values $\boldsymbol{\theta}_j$ from the posterior distribution of $\boldsymbol{\theta}$, we obtain its corresponding *accepted epidemic curves* $\{x_{i,j}\} = \{w_i(G(t;\boldsymbol{\theta}_j))\}$, for weeks $i = 1, \ldots, n = 47$, depicted in Figure 6. All the reported susceptible and confirmed cases are shown by the red and black lines, but only the information in red (up to week $m = 23$) is used to fit the parameters. The blue dashed lines correspond to the 90% predictive intervals constructed from the sampled accepted curves at each week, after week 23. Analysis of the predictive intervals suggests that the curves appear to be very informative before the epidemic peak, but after the peak, the prediction uncertainty tends to be very high.

Using the averaged GDT for Cluster 2, $\{D_i\}$, J different sampled accepted curves $\{x_{i,j}\} = \{w_i(G(t;\boldsymbol{\theta}_j))\}$, and the MR-ARIMA model (3.3) we obtain an ensemble of predictions for weeks $m+1, \ldots, n$, that incorporate information of the posterior distribution of $\boldsymbol{\theta}$ and α. That is, we fit J different MR-ARIMA models using $\boldsymbol{X}^\top = (\boldsymbol{D}, \boldsymbol{x}_j)^\top$ with $\boldsymbol{D} = (D_1, \ldots, D_n)^\top$ and $\boldsymbol{x}_j = (x_{1,j} \ldots, x_{n,j})^\top$.

The model (3.3) with lowest AIC has $p = 1$, $d = 1$ and $q = 0$, and estimates of its coefficients summarized over a set of $J = 370$ fitted models are reported in Table 2.

Figure 7 presents a point forecast and predictive intervals obtained from

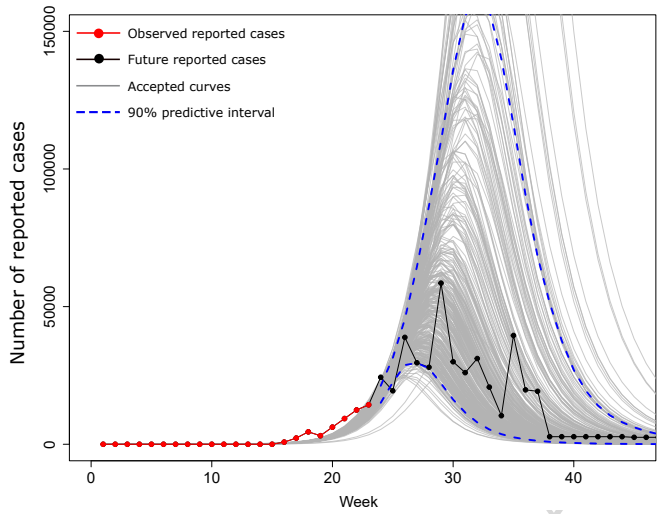

Figure 6: Accepted epidemic curves for the 2014 chikungunya outbreak in the Dominican Republic.

Table 2: Summary of coefficient estimates from 370 models MR-ARIMA(1,1,0)

	AR ($\hat{\phi}_1$)	GDT Cluster 2 (D_i)	Accepted FM ($x_{i,j}$)
1st quartile	−0.1731	16,986	6,897,671
Mean	−0.1648	17,065	7,627,266
3rd quartile	−0.1515	17,280	8,056,531

the fitted ensemble of MR-ARIMAs. The point forecasts correspond to the mean of the ensemble at each week. A predictive interval could also be obtained as quantiles of the ensemble at each week. The resulting bands would incorporate the variability associated with the posterior distribution, but they would not yet include the uncertainty that arises from the MR-ARIMA prediction. To better describe the overall uncertainty, the bands in Figure 7 are obtained as the quantile intervals plus/minus two mean standard MR-ARIMA errors at each week.

We compare the point predictions from MR-ARIMA with predictions obtained by the following three competing methods.

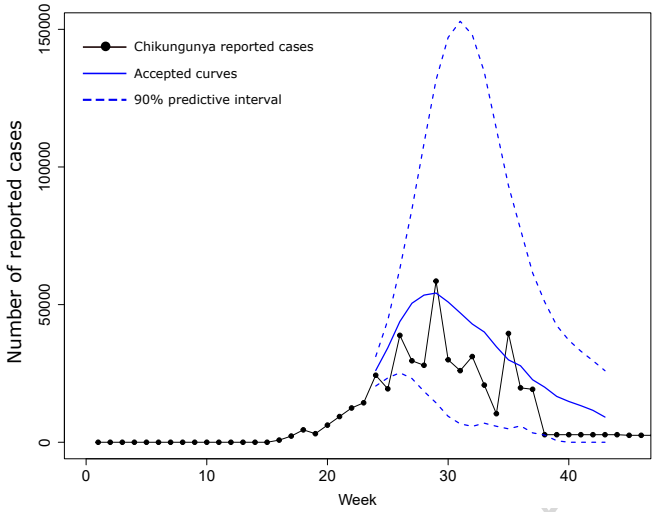

Figure 7: MR-ARIMA forecast of chikungunya in Dominican Republic in 2014, using GDT and accepted curves from the ODE epidemic model.

First, we fit a traditional ARIMA model, without exogenous regressors, for the reported cases from week 16 to 23 by selecting a model that minimizes AIC. The obtained model corresponds to ARIMA(0,1,0) with drift (estimated as 1936.14).

The second model (ODE-MAP) corresponds to a point forecast based on the sampled predictive curves obtained in Figure 6. As the ARIMA model, ODE-MAP is fitted using only on the reported cases and the point parameter estimate that maximizes the *a posteriori* probability (MAP).

The third model (Pois-C2) corresponds to a generalized linear model (GLM) that uses only the covariate D_i (averaged GDT of Cluster 2) that is used in the MR-ARIMA model. In the fitted Poisson regression both the intercept and covariate are statistically significant.

All the produced forecasts are presented in Figure 8 and evaluated based on the mean absolute percentage error (MAPE) and root-mean-square error (RMSE) presented in Table 3.

In spite of the fact that ODE-MAP reports the smallest MAPE, it

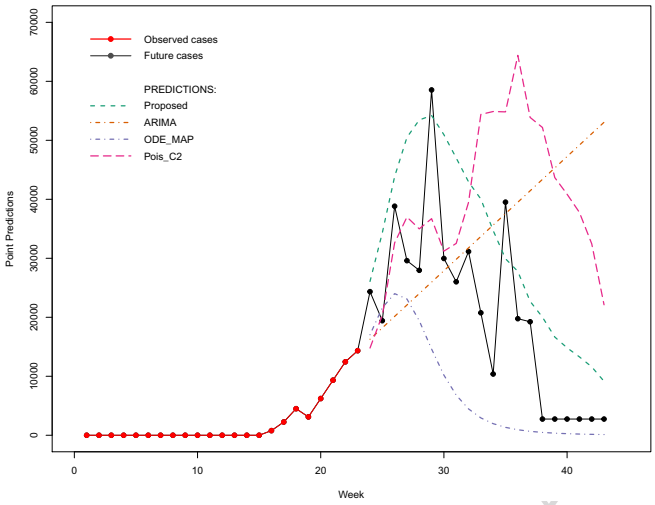

Figure 8: Predictions of chikungunya cases based on the four models.

is clear from Figure 8 that the forecast from the proposed MR-ARIMA model outperforms all competing predictions for peak position and outbreak duration.

Table 3: Performance of the point forecasts from four models

Error measure	Proposed MR-ARIMA	ARIMA	ODE-MAP	Pois-C2
MAPE	1.69	5.33	0.72	4.49
RMSE	14790.76	28046.58	17856.05	27766.66

5 Discussion

We have presented a new multi-source integrative approach for forecasting emerging diseases. In particular, we focus on climate-sensitive vector-borne diseases for which we have observed dynamics in other populations, but have only few records in the evolving outbreak for the area of interest. Using the early weeks' outbreak information, we evaluate the whole epidemic curve

related to the number of new symptomatic cases. Specifically, we estimate when the epidemic will reach its highest point and the outbreak duration.

Based on the available surveillance records, we select auxiliary data that can describe the epidemic evolution, and using this information we also fit an epidemic model. The predictions obtained from both sources are combined, and their contribution is estimated by the parameter fitting process of a regression model with ARMA errors.

We harness GDT since chikungunya is spread by the same principal vectors as dengue, and there is evidence that GDT is able to capture the dengue activity in countries with benign environment to the vector. With this variable, we aim to capture information on the vector population density, but mainly on the human-vector interaction, which is affected by the climate and socioeconomic conditions. Based on the selected GDT information and the fitted epidemic model, we are able to produce an epidemic prediction even under limited official public health data. The point estimate successfully predicts the epidemic peak on week 29 and, along with its predictive bands, it describes a more realistic slower decay, compared to any MCMC sampled epidemic curve.

We base our forecast for reported cases on the local transmissions. Although this seems appropriate for DR, the approach must be modified for countries with high number of international travelers and environmental conditions that have important effect on the local mosquito population.

While the obtained point forecast can capture some important features of the epidemic curve, the prediction intervals are very wide, especially near the epidemic peak. Reduction of the uncertainty can be attempted by using more informative priors or, as it is done in the competing DARPA models, by explicitly introducing more information from additional sources. For instance, additional data could reduce the uncertainty on the under-reporting rate, or could be used to model a dynamic under-reporting. Furthermore, as shown by [22, 33], complementary biosurveillance

information at a multi-scale level can be obtained using methods of topological data analysis. Overall, the epidemic surveillance data problem is highly complex to model and has been addressed very rarely. With the introduction of auxiliary information, e.g., from web searches and topological summaries, we can explore more robust models.

Acknowledgments

This work is supported by National Science Foundation under Grants No. DMS 2027793 and DMS 1925346, and National Aeronautics and Space Administration under Grant No. 80NSSC20K1579.

References

[1] **Appice, A, Gel, Y.R., Iliev, I., Lyubchich, V., and Malerba, D.** (2020). A multi-stage machine learning approach to predict dengue incidence: a case study in Mexico, *IEEE Access*, **8**, 52713-25.

[2] **Bailey, N. T. J.** (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*, London: Charles Griffin & Company Ltd.

[3] **Castro, J., Torres, J., Oletta, J. and Strauss, R.** (2016). Google trend tool as a predictor of chikungunya and Zika epidemic in a environment with little epidemiological data, a Venezuelan case, *International Journal of Infectious Diseases*, **53**, 133–134.

[4] **CDC and PAHO** (2011). *Preparedness and Response for Chikungunya Virus: Introduction in the Americas*, Washington, DC: Pan American Health Organization.

[5] **Chan, E. H., Sahai, V., Conrad, C. and Brownstein, J. S.** (2011). Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance, *PLoS Neglected Tropical Diseases*, **5 (5)**, e1206.

[6] **Christen, J. A. and Fox, C.** (2010). A general purpose sampling algorithm for continuous distributions (the t-walk), *Bayesian Analysis*, **5 (2)**, 263–281.

[7] **Ciampi, A., Appice, A. and Malerba, D.** (2010). Discovering trend-based clusters in spatially distributed data streams, in *International Workshop of Mining Ubiquitous and Social Environments*, 107–122.

[8] **Cotter, S. L., Dashti, M., Robinson, J. C. and Stuart, A. M.** (2009). Bayesian inverse problems for functions and applications to fluid mechanics, *Inverse Problems*, **25**, 115008.

[9] **Dashti, M. and Stuart, A. M.** (2013). The Bayesian approach to inverse problems, *ArXiv e-prints*, 1302.6989.

[10] **De Angelis, D., Presanis, A. M., Birrell, P. J., Tomba, G. S. and House, T.** (2015). Four key challenges in infectious disease modelling using data from multiple sources, *Epidemics*, **10**, 83–87.

[11] **Del Valle, S. Y., McMahon, B. H., Asher, J., Hatchett, R., Lega, J. C., Brown, H. E., Leany, M. E., Pantazis, Y., Roberts, D. J., Moore, S., Peterson, A. T., Escobar, L. E., Qiao, H., Hengartner, N. W. and Mukundan, H.** (2018). Summary results of the 2014-2015 DARPA chikungunya challenge, *BMC Infectious Diseases*, **18 (1)**, 1–14.

[12] **El-Metwally, A. A.** (2015). Google search trend of dengue fever in developing countries in 2013-2014: An internet-based analysis, *Journal of Health Informatics in Developing Countries*, **9 (1)**.

[13] **Escobar, L. E., Qiao, H. and Peterson, A. T.** (2016). Forecasting chikungunya spread in the Americas via data-driven empirical approaches, *Parasites & Vectors*, **9 (1)**, 1–12.

[14] **Gluskin, R. T., Johansson, M. A., Santillana, M. and Brownstein, J. S.** (2014). Evaluation of internet-based dengue query data: Google dengue trends, *PLoS Neglected Tropical Diseases*, **8 (2)**, e2713.

[15] **Huang, X., Iliev, I.R., Lyubchich, V., and Gel, Y.R.** (2018). Riding down the Bay: Spacetime clustering of ecological trends, *Environmetrics*, **29(5-6)**, e2455.

[16] **Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F.** (2020). forecast: Forecasting Functions for Time Series and Linear Models, URL `https://CRAN.R-Bproject.org/package=forecast`, R package version 8.12.

[17] **Johansson, M. A., Powers, A. M., Pesik, N., Cohen, N. J. and Staples, J. E.** (2014). Nowcasting the spread of chikungunya virus in the Americas, *PloS one*, **9 (8)**, e104915.

[18] **Kermack, W. O. and McKendrick, A. G.** (1927). A contribution to the mathematical theory of epidemics, *Proc. R. Soc. Lond. B Biol. Sci.*, **115 (772)**, 700–721.

[19] **Kermack, W. O. and McKendrick, A. G.** (1932). Contributions to the mathematical theory of epidemics. II. The problem of endemicity, *Proc. R. Soc. Lond. A*, **138 (834)**, 55–83.

[20] **Kermack, W. O. and McKendrick, A. G.** (1933). Contributions to the mathematical theory of epidemics. III. – Further studies of the problem of endemicity, *Proc. R. Soc. Lond. A*, **141 (843)**, 94–122.

[21] **Lazer, D., Kennedy, R., King, G. and Vespignani, A.** (2014). The parable of Google Flu: traps in big data analysis, *Science*, **343 (6176)**, 1203–1205.

[22] **Lo, D. and Park, B.** (2018). Modeling the spread of the Zika virus using topological data analysis, *PLOS One*, **13(2)**, p.e0192120.

[23] **Lyubchich, V. and Gel, Y. R.** (2018). funtimes: Functions for Time Series Analysis, `https://CRAN.R-Bproject.org/package=funtimes`, R package ver. 5.0.

[24] **McGough, S. F., Brownstein, J. S., Hawkins, J. B. and Santillana, M.** (2017). Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data, *PLoS Neglected Tropical Diseases*, **11 (1)**, e0005295.

[25] **Milinovich, G. J., Williams, G. M., Clements, A. C. and Hu, W.** (2014). Internet-based surveillance systems for monitoring emerging infectious diseases, *The Lancet Infectious Diseases*, **14 (2)**, 160–168.

[26] **Pacheco, Ó., Martinez, M., Alarcón, Á., Bonilla, M., Caycedo, A., Valbuena, T. and Zabaleta, A.** (2017). Estimation of underreporting of Chikungunya virus infection cases in Girardot, Colombia, from November, 2014, to May, 2015, *Biomédica*, **37 (4)**, 507–515.

[27] **PAHO** (2018). Chikungunya, `https://www.paho.org/hq/index.php?option=`

com_topics&view=article&id=343&Itemid=40931&lang=en, [Online; accessed 2018-10-04].

[28] **Paixão, E. S., Teixeira, M. G. and Rodrigues L. C.** (2018). Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases, *BMJ Global Health*, **3 (Suppl 1)**, e000530.

[29] **R Core Team** (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-Bproject.org/.

[30] **Safta, C., Ray, J., Sargsyan, K., Lefantzi, S., Cheng, K. and Crary, D.** (2011). Real-time characterization of partially observed epidemics using surrogate models, Technical report, Sandia National Laboratories.

[31] **Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O. and Brownstein, J. S.** (2015). Combining search, social media, and traditional data sources to improve influenza surveillance, *PLoS Computational Biology*, **11 (10)**, e1004513.

[32] **Schaeffer, E. D., Testa, J. M., Gel, Y. R. and Lyubchich, V.** (2016): "On information criteria for dynamic spatio-temporal clustering," in Banerjee, A., Ding, W., Dy, J., Lyubchich, V., Rhines, A., Ebert-Uphoff, I., Monteleoni, C. and Nychka, D. eds., *Proceedings of the 6th International Workshop on Climate Informatics: CI 2016*, 5–8, NCAR Technical Note NCAR/TN-529+PROC.

[33] **Soliman, M., Lyubchich, V. and Gel, Y.R.** (2020). Ensemble forecasting of the Zika spacetime spread with topological data analysis, *Environmetrics*, **31(7)**, e2629.

[34] **Strauss, R. A., Castro, J. S., Reintjes, R. and Torres, J. R.** (2017). Google dengue trends: an indicator of epidemic behavior. The

Venezuelan Case, *International Journal of Medical Informatics*, **104**, 26–30.

[35] **Stuart, A. M.** (2010). Inverse problems: A Bayesian perspective, *Acta Numerica*,**19**, 451–559.

[36] **Sullivan, T. J.** (2015). *Introduction to Uncertainty Quantification*, *Texts in Applied Mathematics*, **63**, Cham: Springer.

[37] **Van Bortel, Dorleans, W., F., Rosine, J., Blateau Aand Rousseau, D., Matheus, S., Leparc-Goffart, I., Flusin, O., Prat, C., Césaire, R., Najioullah, F., Ardillon, V., Balleydier, E., Carvalho, L., Lemaitre, A. Noel, H., Servas, V., Six, C., Zurbaran, M., Léon, L., Guinard,A., van den Kerkhof, J., Henry, M., Fanoy, E., Braks, M., Reimerink, J., Swaan, C., Georges, R., Brooks, L., Freedman, J., Sudre, B., and Zeller, H.** (2014). Chikungunya outbreak in the Caribbean region, December 2013 to March 2014, and the significance for Europe, *Eurosurveillance*, **19 (13)**, 20759.

[38] **Yakob, L. and Clements, A. C. A.** (2013). A mathematical model of chikungunya dynamics and control: the major epidemic on Réunion Island, *PloS One*, **8 (3)**, e57448.

[39] **Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N. and Santillana, M.** (2017). Advances in using Internet searches to track dengue, *PLoS Computational Biology*, **13 (17)**, e1005607.

L. Leticia Ramirez Ramirez
CIMAT, Jalisco S/N, Col. Valenciana
CP: 36023. Guanajuato, Gto, Mexico
E-mail: leticia.ramirez@cimat.mx

Vyacheslav Lyubchich
Chesapeake Biological Laboratory, UMCES
146 Williams St., Solomons, Maryland, 20688, USA
E-mail: lyubchich@umces.edu

Yulia R. Gel
University of Texas at Dallas
Richardson, Texas, 75080, USA
E-mail: ygl@utdallas.edu