SCIENCE SET
OPEN ACCESS PUBLISHERS

Research Article

Journal of Clinical and Biomedical Advances

Diagnosing the Outbreak and Spread of the Chikungunya Virus in Central & South America - The DARPA CHIKV Challenge

Leonard J Pietrafesa[1,3*], Tingzhuang Yan[2], Felix Sarubbi[1], Shaowu Bao[1], Treve M Henwood[1], Samuel Bruttomesso[1], Paul T Gayes[1], and Thomas Karl[4]

[1]*Coastal Carolina University*
[2]*Shaw University*
[3] *North Carolina State University*
[4]*Weather & Climate LLC*

[*]Corresponding author: Leonard J Pietrafesa, PhD, Coastal Carolina University.

Submitted: 18 December 2023 **Accepted:** 20 December 2023 **Published:** 27 December 2023

Citation: *Leonard J Pietrafesa, Tingzhuang Yan, Felix Sarubbi, Shaowu Bao, Treve M Henwood, et al. (2023) Diagnosing the Outbreak and Spread of the Chikungunya Virus in Central & South America - The DARPA CHIKV Challenge. J Clin Bio Med Adv 2(4), 01-10.*

Abstract

In 2014, the U.S. Defense Advanced Research Projects Agency (DARPA) issued a "challenge" to forecast the future outbreaks and spreads of the virus Chikungunya Virus (CHIKV) in Central and South America. We took on the challenge of predicting the cumulative number of Pacific Atlantic Health Organization (PAHO) reported Chikungunya or CHIKV cases by South and Central America PAHO country, and by week. Our predicted variables include cumulative numbers of suspected, confirmed and imported cases. To generate these predictions, we developed a methodology that combines statistical and data-driven empirical epidemiological modeling approaches. The results of our model system, the sources of information and a predictive algorithm capability are documented and provided. At the end of the CHIKV Challenge, DARPA announced ten Top 10 Teams, out of 487 that entered the challenge. Our Team was one of the Top Ten Awardees. The challenge period of the CHIKV Challenge only lasted from late August of 2014 until March 2015, just seven months, and thus we believe that the results of many of the DARPA Challenge predictive model systems based on Southern Hemisphere spring to fall conditions were skewed accordingly. Basically, many of the 487 models were simple power law curve fits, including some who finished in the Top 10. Had the context continued for five more months, these curve fitting exercises would have failed as the Southern Hemisphere moved from its fall to winter, and the power law curves would have failed to decrease due to planetary seasonal progressions. Our data based mathematical model approach included related weather factors and seasonal intrinsic modes of variability, which would have continued for more seasons and even years. Those other approaches were mathematically correct but epidemiologically implausible. Unfortunately, the appraiser that DARPA employed to evaluate the mathematical prognostic methodologies was not up to the task. Our approach is both mathematically and epidemiologically generic, and will continue to produce accurate forecasts well into the future in the South and Central Americas. As such, we were honored by DARPA to have been in the Top 10 of the contestants. The cash award was divided up amongst several graduate students. This manuscript documents the successful mathematical methodology that we employed.

Keywords: Chikv, Infectious Diseases

Introduction

The Chikungunya Virus (CHIKV) is spread to people by the bite of an infected mosquito. The most common symptoms of infection are fever and joint pain. Other symptoms may include headache, muscle pain, joint swelling, or rash. Outbreaks have occurred in countries in Africa, Americas, Asia, Europe, and the Caribbean, Indian and Pacific Oceans. There is a risk the virus will be spread to unaffected areas by infected travelers. There is currently no vaccine to prevent or medicine to treat Chikungunya Virus infections. The disease was first detected in the Americas in 2013. Chikungunya usually does not cause death, but the symptoms can be severe and debilitating. The most common symptoms are joint aches and pains. The disease causes fever, fatigue, headaches, muscle pain, skin rashes and depression. In 2014, to accelerate the development of new infectious disease forecasting methods, DARPA launched its CHIKV Challenge. It was intended to address the further outbreak and spread of the virus and also to seek new approaches for other viruses in-kind. It was a laudable preemptive move on the part of DARPA. There was no upfront funding but the brass ring was the offer of cash prizes to the Top 10 finishers, to have been determined by an organization hired by DARPA to assess the goodness of the

forecasts. Four hundred and eighty-seven Teams competed in the contest. The Top 10 Teams received fully paid travel expenses to the DARPA Awards ceremony, held in May 2015, at the DARPA Headquarters Non-Secure facility in Washington DC. This manuscript documents and summarizes the efforts of our Top 10 Team.

Methodology Description

We first pre-processed the weekly updated numbers of suspected, confirmed and imported Chikungunya cases in Central and South American countries/territories provided by PAHO to ensure data quality and consistency. For example, if the cumulative case numbers of the most recent week were less than the numbers of the previous week, we were forced to revisit the reported numbers and the trends of the past several weeks' cumulative numbers of suspected cases, confirmed cases and imported cases to identify which data were more reasonably correct. Based on the processed PAHO data, we categorized each of the 49 countries into two groups: a) countries with only recent CHIKV transmission records, defined as < 25 weeks of autochthonous cases were classified into Group 1; and b) countries with an extended period of CHIKV transmission, defined as equal to or greater than (=/>) 25 weeks of autochthonous, i.e. native or indigenous, cases were classified into Group 2. Initially, 42 countries were in Group 1, while only 7 countries were in Group 2. As time progressed, more countries moved from Group 1 into Group 2, in keeping with the classification criteria. As a part of the software package we provide within, a program is designed to automatically check the number of non-zero record weeks for each country.

Here we differentiated between Groups 1 and 2 by looking at the apparent internal modes of variability that are buried within the disease data sets themselves; as described below in Section 3 of the text. Basically, we subjected the data to several different mathematical transforms which offered the potential for revealing the periods of time over which there are significant, well defined, energetic modes of variability that are elemental, innate, and constitutive to the individual and collective data sets. What we found was that the data sets all contained robust bi-weekly to tri-weekly to monthly signals but also to bi-monthly and multi-monthly signals. Thus, as the time series began and were thus relatively short on the front end, the time series were quite limited and compromised our ability to project the time series into the future with a high confidence of certainty. The ability to accurately predict the future from the past and present was limited to about 10% of the total period of the individual time series. So, to 0th (zeroth) order we separated the collective time series into those that had data for less than 25 weeks and those that had data for more than 25 weeks. At more than 25 weeks the variability that is inherent to CHIKV time series seems to have been for the most part realized; given the nominal 7-month length of this scoping study. This will be re-discussed further below (in Section 3) with schematic proof of the seemingly subjective, but actually quasi-objective choice that we made based on our initial assessments.

a. Forecasting for Group 1 Countries

The model system first reads in the pre-processed PAHO input cumulative number of cases, analyzes the distribution pattern of the short period data objectively and then automatically constructs either a linear (least squares) or a polynomial regression model that best fits the past data with a power function. This constructed regression model was then used for projections.

b. Forecasting for Group 2 Countries

PAHO-reported cumulative case number includes autochthonous transmission case numbers and imported case numbers. For each Group 2 country, weekly incidence time series were extracted from cumulative number and used for key parameter identification. The model first compared the cumulative autochthonous transmission case (ATC) number to imported case (IC) numbers of the most recent week. The following three cases were considered:

i) If the ATC was ten or more times larger than TC, then the weekly autochthonous transmission case number was used to identify key parameters critical for attempting to accurately predict disease spread within a country, and these parameter estimates, along with various other data sources detailed below, were used to drive Group 2 country-specific epidemiological CHIKV transmission models, yielding incidence projections, and then used to map incidence projections into PAHO-reported case projections.

ii) If the IC was ten or more times larger than the ATC, the weekly imported case number was used to identify key parameters for the imported case prediction model, and then it was mapped to PAHO reported case projections.

iii) If the ratio of ATC/IC was between 0.1~10, the weekly autochthonous case and imported case time series were used to identify two separate sets of key parameters, and into the autochthonous cases were classified into Group 2. Initially, 42 countries were in Group 1, while only 7 countries were in Group 2. As time progressed, more countries moved from Group 1 into Group 2, in keeping with the classification criteria. As a part of the software package we provide within, a program is designed to automatically check the number of non-zero record weeks for each country.

Here we differentiated between Groups 1 and 2 by looking at the apparent internal modes of variability that are buried within the disease data sets themselves; as described below in Section 3 of the text. Basically, we subjected the data to several different mathematical transforms which offered the potential for revealing the periods of time over which there are significant, well defined, energetic modes of variability that are elemental, innate, and constitutive to the individual and collective data sets. What we found was that the data sets all contained robust bi-weekly to tri-weekly to monthly signals but also to bi-monthly and multi-monthly signals. Thus, as the time series began and were thus relatively short on the front end, the time series were quite limited and compromised our ability to project the time series into the future with a high confidence of certainty. The ability to accurately predict the future from the past and present was limited to about 10% of the total period of the individual time series. So, to 0th (zeroth) order we separated the collective time series into those that had data for less than 25 weeks and those that had data for more than 25 weeks. At more than 25 weeks the variability that is inherent to CHIKV time series seems to have been for the most part realized; given the nominal 7-month length of this scoping study. This will be re-discussed further below (in Section 3) with schematic proof of the seemingly subjective, but actually quasi-objective choice that we made based on our initial assessments.

c. Forecasting for Group 1 Countries

The model system first reads in the pre-processed PAHO input cumulative number of cases, analyzes the distribution pattern of the short period data objectively and then automatically constructs either a linear (least squares) or a polynomial regression model that best fits the past data with a power function. This constructed regression model was then used for projections.

d. Forecasting for Group 2 Countries

PAHO-reported cumulative case number includes autochthonous transmission case numbers and imported case numbers. For each Group 2 country, weekly incidence time series were extracted from cumulative number and used for key parameter identification. The model first compared the cumulative autochthonous transmission case (ATC) number to imported case (IC) numbers of the most recent week. The following three cases were considered:

i) If the ATC was ten or more times larger than TC, then the weekly autochthonous transmission case number was used to identify key parameters critical for attempting to accurately predict disease spread within a country, and these parameter estimates, along with various other data sources detailed below, were used to drive Group 2 country-specific epidemiological CHIKV transmission models, yielding incidence projections, and then used to map incidence projections into PAHO-reported case projections.

ii) If the IC was ten or more times larger than the ATC, the weekly imported case number was used to identify key parameters for the imported case prediction model, and then it was mapped to PAHO reported case projections.

If the ratio of ATC/IC was between 0.1~10, the weekly autochthonous case and imported case time series were used to identify two separate sets of key parameters, and into the construct of two projection models. The PAHO-reported case projection became the sum of the above two (i + ii) model projections. We note here that Poisson Regression in a log linear model was applied based on available key parameters for the above cases. The model extends the traditional linear model, which estimates the parameters of the model numerically through an iterative fitting process. The methodology allows simultaneous testing and modeling of multiple independent variables that provide revealing insights into the relationships between several independent potential predictor variables and the dependent variable. Cross-correlations between cumulative cases and key epidemiological parameters were then calculated and analyzed. Examples of this process and methodology are given in the description and discussion parts of the text to follow.

Data Sources

Below we provide a description of which data were used to create prediction models and how and why the data sources were selected. For the data sources used to make our predictions, we address the question of representativeness of the data sources and if not, we mention which groups were underrepresented in the data and what were the potential impacts caused by these exclusions. We also describe the accessibility and affordability of the data sources.

To predict PAHO-reported cumulative case numbers for Group 1 and Group 2 countries, we considered a number of distinct, complementary data sources: epidemiological data (PAHO and MOH CHIKV data), human population data, mosquito data, environmental (country-wide) temperature and precipitation, geological terrain and socioeconomic data. All of those data are potentially associated with either ATC and or IC numbers. Temperature and precipitation data, along with entomological data, were obtained and processed for cross-correlation analysis with the weekly disease data. Data sources that we utilized in our model scheme, include:

1. DARPA CHIKV: https://www.innocentive.com/darpa-chikv-challenge-resources
2. PAHO: Weekly updated numbers of suspected and confirmed Chikungunya cases in PAHO countries/territories: http://www.paho.org/hq/index.php?option=com_topics&view=rdmore&tt=PAHO%2FWHO+Data%2C+M aps+and+Statistics&id=5927.
3. Chikungunya genomic information: European Virus Archive: Complete CHIKV phylogeny, PCR systems, amino acid, and nucleotide sequence of St. Martin strain: http://www.european-virus- archive.com/article147.html. Chikungunya virus strain S27-African prototype, complete genome-GenBank Accession #: AF369024. Chikungunya virus strain La Reunion strain, complete genome-GenBank Accession #: DQ443544.2
4. IATA: International Air Transport Association. Commercial flight database. IATA is the IATA is the trade association for the world's airlines.
5. Bureau of Transportation Statistics, Research and Innovative Technology Administration: http://www.rita.dot.gov/bts/press_releases/bts024_14. http://www.transtats.bts.gov/Fields.asp?Table_ID=260
6. Centers for Disease Control and Prevention Resources: http://www.cdc.gov/chikungunya/resources/index.html. Division of Vector-Borne Diseases
7. Census data. IPUMS-International -World's largest collection of publicly available individual-level census data. Integrates samples from population censuses from around the world taken since 1960. For the U.S., IPUMS-USA, which is optimized for U.S. research. https://usa.ipums.org/usa/. https://international.ipums.org/international/
8. Climate and weather data from the NOAA National Climatic Data Center: http://www.ncdc.noaa.gov/
9. Climate and weather data -Climate Prediction Center (NOAA) links. Main page: http://www.cpc.ncep.noaa.gov/products/african_desk/cpc_intl/index.shtml. Weather page for Caribbean - Central America: http://www.cpc.ncep.noaa.gov/products/african_desk/cpc_intl/camerica/camerica.shtml Climate forecasting page including Caribbean -Central America: http://www.cpc.ncep.noaa.gov/products/international/nmme/nmme1.shtml
10. Vector data. Ecological distribution. Feeding behavior: day/night, locations. Weather channel's mosquito activity forecast: http://www.weather.com/activities/homeandgarden/home/mosquito/index.html Mosquito- based arbovirus surveillance software. Example: CDC for West Nile virus: http://www.cdc.gov/westnile/resourcepages/mosqSurvSoft.html
11. Weather and Climate forecast model data. http://iridl.ldeo.columbia.edu/maproom/Health/Regional/Africa/Malaria/IRI_Seasonal_Precip.html

For the CHIKV Forecasts of Group 2 countries in particular, we considered all of the above 11 data sets, which are representative of what has occurred epidemiologically, and what the contemporaneous basis environmental and other conditions were. Mosquito population size and bite rates, which are crucial key factors to Chikungunya autochthonous transmission, are a function of climatic factors, such as temperature and rainfall, which also affect mosquito behavior. Specifically, we attempted to integrate those data to parameterize this dependency in the ATC prediction model. However, given the relatively short period lengths of the country data sets, the lack of complete country-by-country environmental data sets and the relative shortness of the period of performance of this exploratory study, which only covered southern hemisphere spring to summer, we could not fully utilize the various data sets in our relatively compact study effort. Nonetheless, we believe that we have made significant progress in the development of reliable CHIKV forecasting tools. The data sources listed above are representative, and no groups are underrepresented, to our knowledge; so, there are no potential impacts other than to point out that with sufficiently longer data sets a far more complete model system is possible. The data were all open to the public; therefore, they are accessible and affordable.

Below we present examples of data that we collected and include:

a) a representative sample of the PAHO data sets (Table 1)
b) a representative sample of the observed air temperature and precipitation data sets (Table 1) plus, a plot including both air temperature and precipitation (Figure 1)
c) two representative examples of a representative numerical model-based forecast (in the case shown, model CFSv2) of environmental state variables, air temperature and precipitation, respectively, for the first three months of 2015 (Figures 2 and 3).

Table 1: Rochambeau French Guiana (Representative Example) Data by Week.

Year	week	Avg Temp (deg),	Total Precip (inches)	Cases
2014	1	26.24	0.33	0
2014	2	21.45	2.1	0
2014	3	24.24	1.33	0
2014	4	21.33	5.3	1
2014	5	23.51	4.96	1
2014	6	26.4	1.73	1
2014	7	26.61	0.85	1
2014	8	25.58	2.41	2
2014	9	26.37	0.16	10
2014	10	26.19	0.09	5
2014	11	26.78	0.28	9
2014	12	27.15	0.6	1
2014	13	26.55	1.09	3
2014	14	27.15	0.15	1
2014	15	27.07	1.12	6
2014	16	27.11	0.89	4
2014	17	27.44	0.13	6
2014	18	26.94	0.47	13
2014	19	26.23	0.76	16
2014	20	27.07	3.71	41
2014	21	26.76	1.5	54
2014	22	27.14	3.16	46
2014	23	25.96	1.35	24
2014	24	26.11	1.25	72
2014	25	26.38	2.06	70
2014	26	26.08	2.03	71
2014	27	25.8	2.61	71
2014	28	26.39	0.54	71
2014	29	25.79	0.03	139
2014	30	25.62	0.05	141
2014	31	26.33	2.08	202
2014	32	25.51	0.87	200
2014	33	26.27	3.44	471

2014	34	26.59	0.46	452
2014	35	26.42	0.72	0
2014	36	26.01	0.39	3350
2014	37	27.31	0.24	0
2014	38	27.01	1.82	612
2014	39	27.12	0.06	2445
2014	40	27.19	0.09	0
2014	41	26.96	0.06	0
2014	42	26.89	0.03	1050
2014	43	27.19	0.02	0
2014	44	27.03	0.05	1834
2014	45	26.77	0.06	0
2014	46	26.79	1.54	0
2014	47	26.71	0.6	1373
2014	48	26.39	0.04	322
2014	49	26.24	1.25	0
2014	50	26.37	1.8	878
2014	51	26.35	1.07	0
2015	1	25.89	0.54	0
2015	2	26.41	1.48	0
2015	3	25.99	2.32	0
2015	4	25.81	2.14	0
2015	5	26.43	0.83	2209

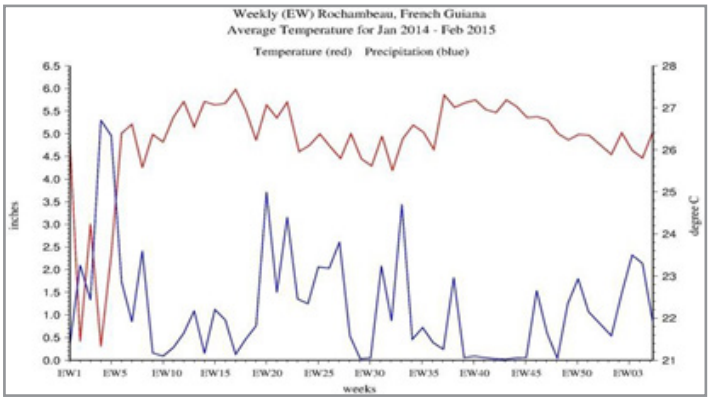

Figure 1: Weekly Rochambeau, French Guiana Average Temperature (red) and Precipitation (blue).

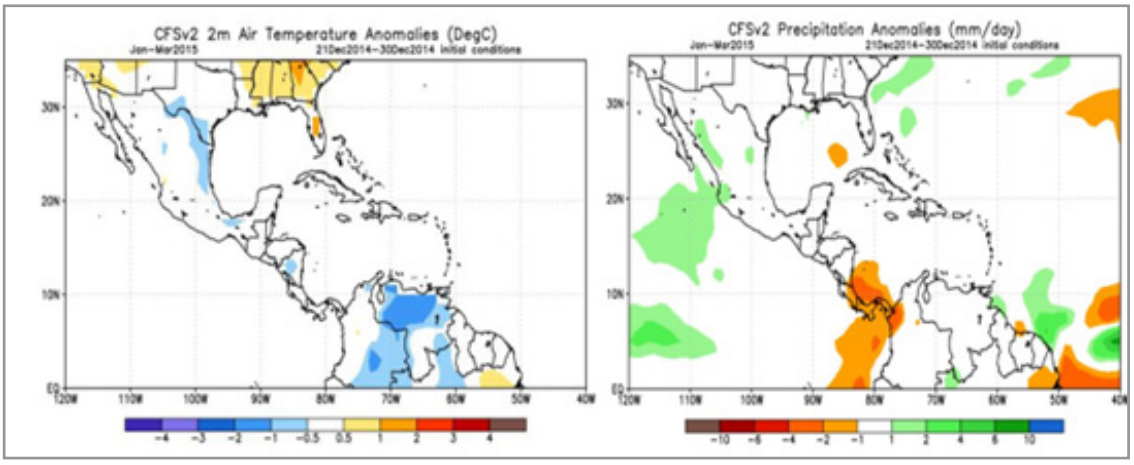

Figure 2: Numerical model CFSv2 based Forecast of precipitation anomalies

Figure 3: Numerical model CFSv2 Forecast of air temperature anomalies

Model Robustness

Here we describe and evaluate our model robustness by describing our model components more completely than was presented in the Introduction of this report, and then by testing our model prognostic output against actual observations. Further, we have provided several walk-throughs of how well our initial model can perform when utilized by an independent assessor. We provide a clear description of assumptions made in the development of the model and we explain how sensitive the model is to certain inputs i.e. perturbations in the data. In developing our response to the DCC, we employed a combination of stochastic, deterministic and empirical mathematical methodologies in a unique manner that allowed us: to conduct diagnostics of non-linear (NL) and non-stationary (NS) data sets; to analyze the data sets, utilize other existing environmental, ecological data sets and weather and climate data sets and numerical model output, entomological information and socio-economic considerations; and to make forecasts of CHIKV. As the data were expected to be NS and NL, we employed mathematical methodologies described below to reveal internal modes of temporal and spatial variability buried within the CHIKV data sets. Then we investigated the feasibility of employing suites of mathematical functions along with Cross-Correlations (CCs) of the disease data and ecological, environmental, unintentional shipping of carriers, human vector carriers traveling into and out from affected areas to further elucidate the structure of the data sets. We have developed a mathematical methodology, which rests on a potential combination of Principal Component Analysis (PCA), Empirical Orthogonal Function (EOF) analysis, Fourier (FA), Morlet-Wavelet (M-W) and Hilbert Transform (HT) and Ensemble Empirical Mode Decomposition (EEMD) analyses and identifies mathematical functions to apply to the highly NS and NL data sets [1-4].

We note as an aside here that just as in U.S. National Weather Service Weather forecasting, we could have adopted a deterministic numerical modeling approach, in addition to our statistical and empirical approaches, but there would have had to have been multiple models, some of which are interactively coupled, and while our collective team possesses the ability to conduct those modeling efforts, such an effort was deemed to be well beyond the scope of the present scoping study. Generally, NL and NS disease and environmental data sets contain overall series length trends. We found that the overall bent of the data in the disease time series, being internal and intrinsic to the data set, can only be determined via the employment of a method that is adaptive across the entirety of the data set, from beginning to end, and is the gravest or lowest internal mode buried within the data set. As such, the lowest internal mode can have either no or at most one inflection point. So, it can either go: up in slope; or down in slope; or up and down in slope; or down and up in slope. This definition of trend presumes the existence of a natural time scale to the entirety of the data set time series. All these requirements suggest the adoption of the HT to deal with the issue of the establishment of an overall trend of a time series no matter the degrees of non-stationarity nor non-linearity.

PCA and EOF are statistical procedures that use orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called PCs. This transformation is defined in such a way that the first PC has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. PCs are guaranteed to be independent if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables. EOF analysis is a decomposition of a signal or data set in terms of orthogonal basis functions which are determined from the data. It is the same as performing a PCA on the data, except that the EOF method reveals both time series and spatial patterns. The method of EOF is similar in spirit to harmonic analysis, but harmonic analysis typically employs predetermined orthogonal functions, at fixed frequencies. In some cases the two methods may yield essentially the same results. The basic functions are typically found by computing the eigenvectors of the covariance matrix of the data set. A more advanced technique is to form a kernel out of the data, using a fixed kernel. It is of note here that these products have the potential of revealing the spatial distribution of the percent of variance of the disease that is contained in the family of EOF modes across South and Central America and the SE U.S.A., but that is well beyond the scope of this study.

In time series analysis, the CC between two time series describes the normalized cross covariance function. If we have a pair of stochastic processes that are jointly wide-sense stationary, then the CC of a pair of jointly wide-sense stationary stochastic process can be estimated by averaging the product of samples measured from one process and samples measured from the other (and its time shifts). The samples included in the average can be an arbitrary subset of all the samples in the time series. For a large number of samples, the average converges to the true CC. We assessed correlations of CHIKV with environmental and other factors using CCs. A discussion proceeds and follows the Figures 4 upper and lower panels. In Figure 4 upper and lower respectively, representative CC matrices between CHIKV and Air Temperature and precipitation in French Guiana are presented. What they reveal is that both CHIKV and air temperature and CHIKV and precipitation are highly and significantly correlated (correlation coefficient of 0.4 and 0.5 for CHIKV vs. temperature and CHIKV vs. precipitation, respectively, above the 95% confidence level) but only after the periods of 20-21 weeks. The CFL rises because the more lags, the less sample data points were used in calculating the coefficient. Therefore, with less overlaying data points, the two-time series need to be more "correlated" in order to qualify for 95% CFL. The correlation between the disease breakout and Air Temperature rises rapidly up to about weeks 20-21 and then increases barely and monotonically thereafter to the end of the time series. Thus, the mosquito community appears to respond on queue to increasing air temperature with a lag period from egg and larval stages into adult hood and then to find and come into contact with human beings. Likewise, the CHIKV responds to precipitation over the same 22-23-week period and then remains high but quite spotty. This may reflect the patchy nature of precipitation and the availability of pools of water, etc. for the mosquito population to grow and multiply and reach adulthood. So, the biology, while responding to environmental queues, has its own time scale to work through and as it evolves it must overlap with that intrinsic to the contact of mosquitoes with humans. This is clearly an area of multi-expertise research, including entomologists, epidemiologists, population dynamists, weather and climate experts, socio-economists, demographic and human impacts experts.

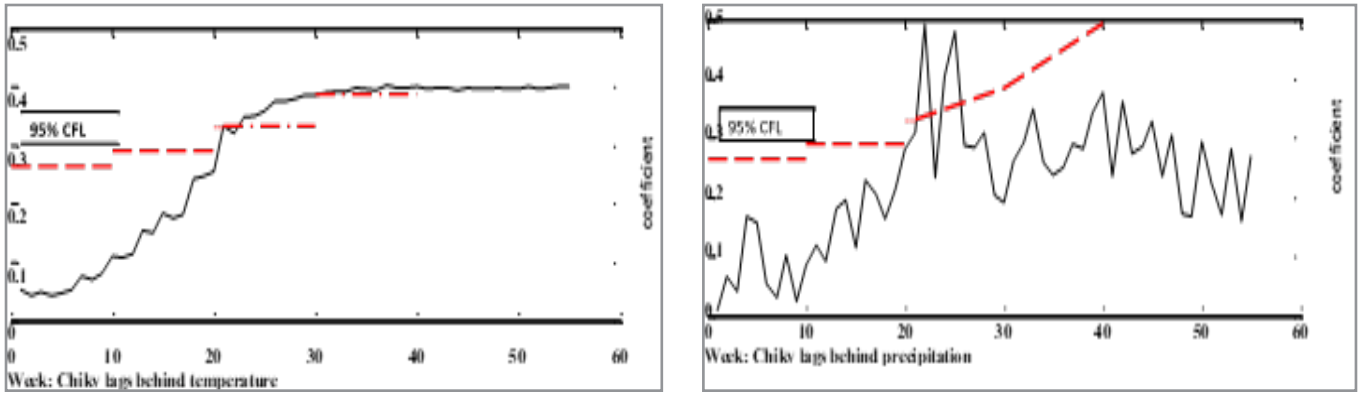

Figure 4: Cross correction between temperatures vs. CHIKV counts (upper panel) and between precipitations vs. CHIKV counts (lower panel).

Employment of the HT is a way to decompose a signal into its hidden, buried, internal modes of variability and to obtain instantaneous frequency data. It is designed to work well for data that is NS and NL. The HT can decompose a signal into functions that are generally simple oscillatory modes as a counterpart to simple harmonic functions. This decomposition method operating in the time domain is adaptive. Since the decomposition is based on the local characteristic time scale of the data, an internal mode can have variable, modulated amplitude and frequency along the time axis. The components of the modes are physically meaningful, for the characteristic scales are defined by the physical data. The decomposition is equivalent to a dyadic filter bank. The original data can be expressed as the real part of the total transform.

In Figures 5a and 5b, the representative results of HT and EEMD decompositions of Rochambeau French Guiana Air Temperatures and Precipitation are presented, in addition to HTs of CHICKV case data in French Guiana (Figure 6) and Martinique (Figure 7). We see that there are five internal modes of buried variability within the temperature and precipitation (Figures 5a, 5b, respectively), and CHIKV data sets (Figure 6 and Figure 7) (in the stacked Air Temperature and Precipitation plots the Modes 5, the trends, are drawn as the red lines through the time series (top panels)). They include (1) bi-weekly, (2) monthly, (3) bi-monthly, (4) 3-monthly and record length trends (red line). The important message here is that the system is one-way coupled and CHIKV is the affected variable, while the others are the influencing, causal factors. We see that for air temperature and precipitation, all internal modes of variability tend to be of equal importance so there is no favorable time period in the causal environmental agents.

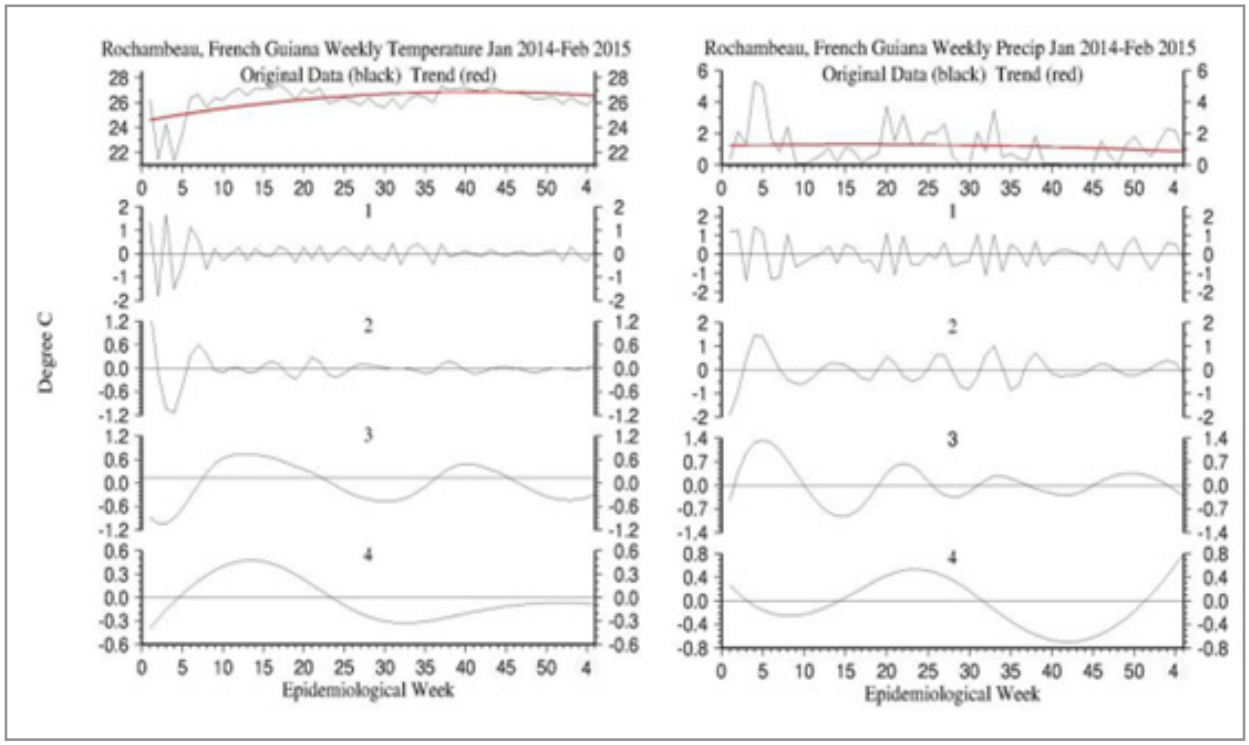

Figure 5: Weekly averaged air temperatures (a) the left panel, and precipitation (b) right panel, and their EEMD IMF decompositions. The top plots are the time series, with IMFs 1, 2, 3 and 4 stacked below and the time series trends are the red lines.

For Group 2 country forecasts, model hind-casting tests demonstrated robust and good agreement with the observations. We used the "Jack-knife kick-off-two methodology" in which we left out two observations at a time from the data set to estimate the bias and standard error and then evaluated the model sensitivity to input data and its forecast uncertainty [1]. Benchmark tests were done by running the model simply using linear regressions without key factors. Model skill was measured using root square error by evaluating model output against the benchmark test output. The model skill was found to be, overall, 18.6% over the benchmark skill, which means that our model system is a considerable improvement in our Group 2 country forecast model even without climate and other factors being uniformly considered; as they were either unavailable for particular countries or because the time series are relatively short and do not allow for sufficient phase lags to have developed. Still, we maintain that in the broader scope of disease forecasting, an underlying model assumption is that the projection variation can be explained by its associations with associated key factors. In the French Guiana HT plots, for the upper three panels and the lower left and center panes, the red stars are the actual observations versus our blue line forecast. However, in the lower right panel of the plot, the blue stars are the actual observations, the black line is our forecast projection and the red dashed line is included to show what a linear projection would result in versus our NL, NS model forecast.

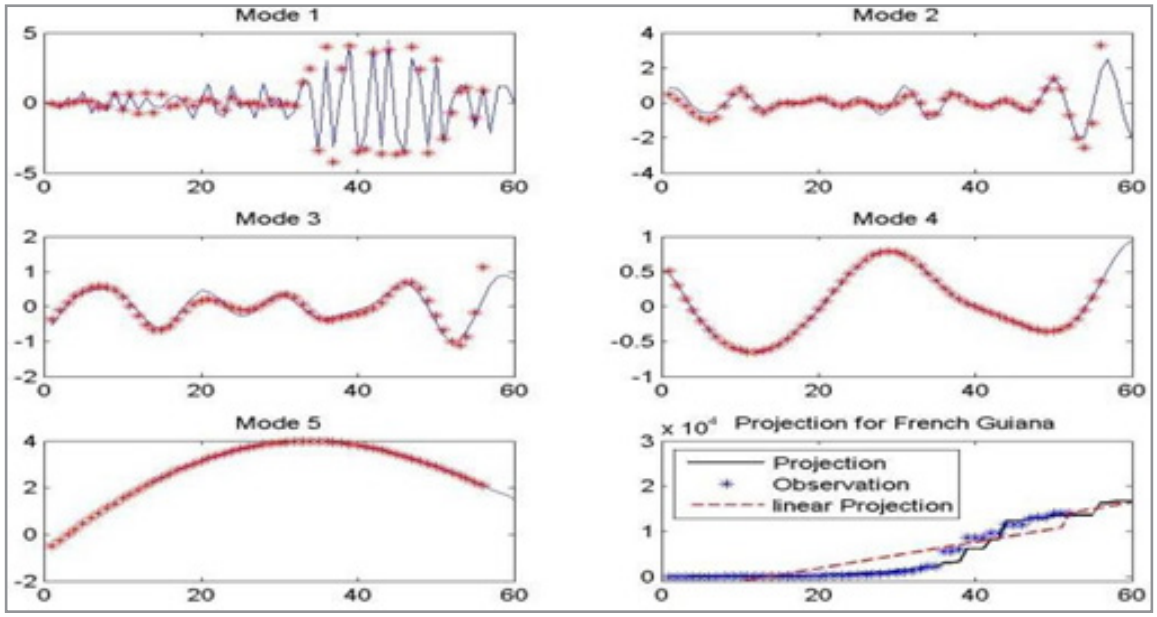

Figure 6: HT decompositions of CHIKV case data in French Guiana.

In the Martinique plots (Figure 7), the red stars are the actual observations and the blue lines are our model projections; including the final projection in the lower right panel.

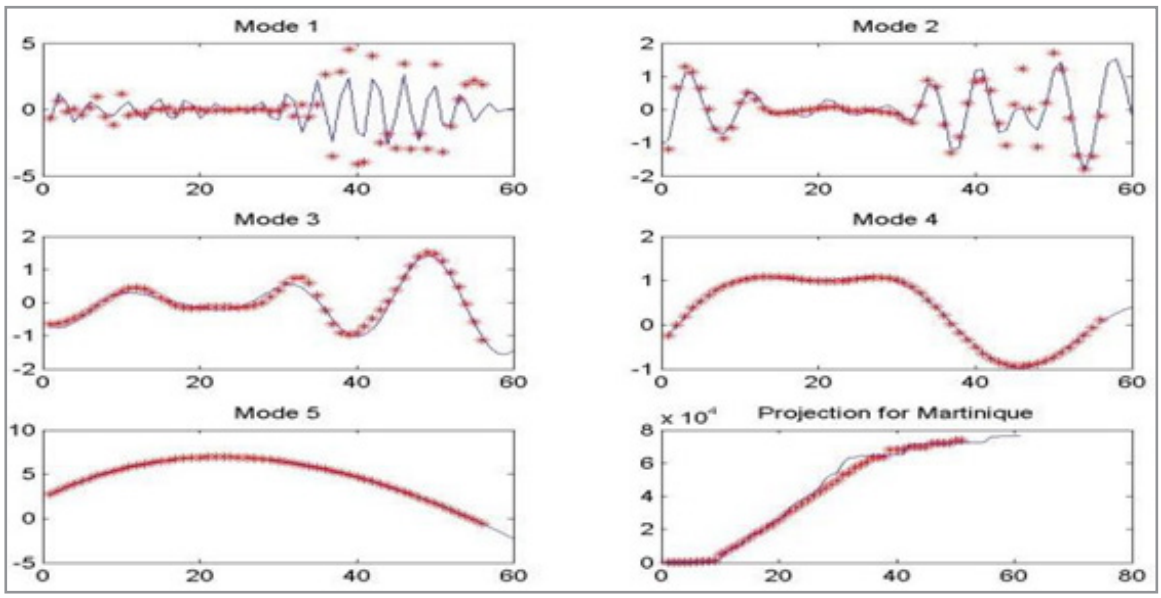

Figure 7: HT decompositions of CHICKV case data in Martinique.

As we described in the text above, we then took the information collectively provided above, extended the internal modes of variability IMFs), including ensemble averaged modulated periods and amplitudes of the same and the extended them into the future to constitute our respective Group 2 CHIKV forecasts. Examples of our forecasts vs. actual observations are next presented for four countries, by way of example (Figure 8). As we can see, the model system is robust in its degree of correctness and repeatability in application from country to country. While not exact, for the most part, the forecasts (blue lines) and the actual outcome, as determined by the observations (red lines) are distinguished by the thickness of the lines. Our Forecast of total cumulative cases of CHIKV for February 2015 for the first six countries, alphabetically speaking, is presented in the abbreviated table below; by way of example (Table 2).

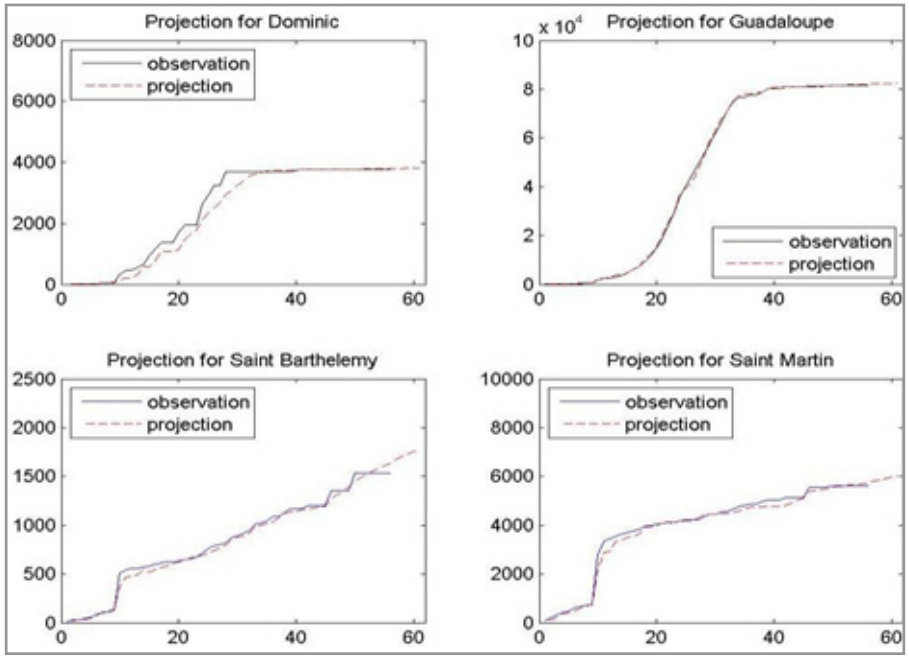

Figure 8: Our predictions vs. Actual Observations of Documented Cases of CHIKV in four (of 49) Central and South American countries

Table 2: Forecast of total cumulative cases of CHIKV for February 2015 for the first six of the forty-nine countries, in alphabetical context of country names, by week, and by way of example

Countries/Municipalities	2-6-2015	2-13-2015	2-20-2015	2-27-2015
Anguilla	112	113	115	116
Antigua & Barbuda	1592	1628	1664	1699
Argentina	47	47	48	49
Aruba	968	980	991	1003
Bahamas	102	104	107	109
Barbados	1906	1940	1975	2010

Applicability

Below we address the question regarding whether or not our methodology and model could be easily applied to additional vector borne diseases (our examples include Dengue Fever, the West Nile Virus and others). Further, we comment upon whether or not our model could inform a mitigation strategy. Our methodology can automatically and robustly identify the patterns hidden in the input data and lead to the construction of models based on these patterns; therefore, it is applicable to a plethora of additional vectored and other diseases, and can easily be applied to other vector borne disease projections, given that the data representing the key factors are available. Here, an emerging school of thought has been to examine the concept of community resilience in the face of routine versus non-routine infectious disease crises.

A routine infectious disease related crisis would include an infection that is endemic, such as cholera in India and Bangladesh, Lyme disease in Connecticut and adjacent areas, malaria in portions of Africa, Tick-borne Spotted Fever in North Carolina, histoplasmosis in the Mississippi River Valley, yearly influenza virus infections, Dengue in the Caribbean and coccidioidomycosis, infections in the Southwest US. Climatic, ecological, weather and vector related factors could play a role in each of these occurrences. A non-routine crisis may be defined as an infectious disease event that catches a community by surprise, where the disruptive impact is greatly enhanced when experts are unable to provide identification of the causative agent and countermeasures are perceived to be ineffective. Examples of non-routine events would include the appearances of the Dengue

virus, Cholera in Bangladesh and Haiti, the occurrence of unusual influenza viruses in Southeast Asia, and the documentation of a unique, Cryptococci fungal infection in the U.S. Pacific Northwest. Climatic, ecological, weather and vector related factors appear to play a role in each of these occurrences. Therefore, our model and its potential applications are certainly not limited to West Nile Virus and Dengue Fever.

This model can inform a mitigation strategy since the domestic transmission and imported cases were separated in the Group 2 projection. We believe that our forecasts are robust and skillful, and are based on relationships, correlative and causal, that the medical and public health communities will be able to take preventative measures to reduce the potential impacts. Additionally, the model can accept diverse sets of data with minimum revision of settings and parameters.

Portability

Our model development approach was and based upon our underlying belief in the principle that electronic health data offers opportunities for analysis leading to the creation and production of disease signature libraries for purposes of epidemiological prognostic forecasting. Diagnosing disease signature patterns offers the potential for proactive health intervention. It offers the prospect of enhancing recognition of both reoccurring routine disease activity and unusual, non- routine disease activity. Forecasting disease signature patterns to enhance pre-emptive warning are crucial components of epidemiological watches and warnings. To accomplish this, data bases necessary for exploitation must be identified, harvested and exploited.

Representation

We have shown and clearly presented examples of the data that we utilized and examples of the results. Our model outputs are suitable for use as a decision support tool. We have provided a clear and accurate visual display of the model output. Our model is easy to run on personal computers. Its output and its accurate visual display are clear, straight forward and easy to be understood by people with and without epidemiological modeling expertise.

Computational Requirements

Our model system requires reduced computational requirements; as compared to a high-performance multi-processor computational platform. Conventional laptops and or desktop computers connected to the Internet at 50 Megs will suffice. The algorithms used in our model system can be run within an approximate 30-minute time frame on a workstation or a laptop in a Windows, Mac or Linux environment with Matlab and SAS software installed. Our model system is scalable to accept diverse sets of data; as needed and appropriate.

Acknowledgements

Coastal Carolina University's Gupta College of Science and North Carolina State University's College of Sciences supported this research via the availability of their respective computational facilities.

References

1. Bendat JS, Piersol AG (2010) Random Data: Analysis and Measurement Procedures. Wiley Publication Corp. Series in Probability and Statistics. DOI:10.1002/9781118032428.
2. Gabor D (1946) Theory of communication. J. IEEE 93: 429-441.
3. Huang NE, Shen Z, Long SR, Wu MC, Shih EH, et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. Roy. Soc. Lond 454: 903-995.
4. Wu Z, Huang NE (2009) Ensemble Empirical Mode Decomposition: a noise assisted data analysis method. Advances in Adaptive Data Analysis 1: 1-41.