Journal of the University of Ruhuna
Volume 9 No 2, December, 2021. Pp 72-85
DOI: http://doi.org/10.4038/jur.v9i2.7977
ISSN 2659-2053

RESEARCH ARTICLE

Identifying Extreme Behaviour and Fitting Empirical Models for Dengue Incidents of Selected Regions in Sri Lanka

S.M.J.H Nisansala*, P. Wijekoon
Department of Statistics and Computer Science
University of Peradeniya, Peradeniya, Sri Lanka.

Abstract: Dengue fever is one of the most significant mosquito-borne diseases caused by a virus. Numerous methods available to predict dengue incidents are mainly focused on the mean features of events. However, understanding the extreme behaviour of dengue incidents is important, and that will allow sufficient time to take the necessary decisions and actions to safeguard the situation for local authorities. Therefore, this study mainly focuses to model the risk of rare dengue events, that is, extreme dengue events, and to identify the best-fitted distributions for the study areas. Further, the weather-based dengue empirical models for dengue incidents were fitted using climatological factors to forecast potential outbreaks. The weekly dengue incidents and climatology data (rainfall, temperature, and relative humidity) from January 2010 to December 2018 for seven administrative districts were collected from the Epidemiology Unit of the Ministry of Health (MoH), and the Meteorology Department of Sri Lanka, respectively. The Extreme value theory (EVT) was used to analyse the extreme dengue incidents, and the negative binomial generalized linear model was used to fit weather-based dengue empirical models. Various lag times between dengue and weather variables were analysed to identify the optimal dengue forecasting period. The best fitted empirical models for dengue incidents were identified for the selected districts. The Generalized Linear Negative Binomial (GLNB) models with monsoon season as a covariate, lag 0 model is the suitable model for Colombo and Gampaha districts, and lag 1 model is the suitable for Kurunegala whereas lag 2 model is the best for Anuradhapura with highest prediction accuracy. For Badulla district, lag 2 model without having monsoon season as a covariate shows highest prediction accuracy. The prediction accuracy is the same for the models with or without having the monsoon season as a covariate for Kandy (lag 2) and Ratnapura (lag 3) districts.

Keywords*: Generalized extreme value distribution, Generalized Pareto distribution, Dengue incidents, Generalized linear negative binomial model, Climatic factors*

Introduction

Dengue fever is that the more rashly spreading mosquito-borne viral infection that appears in tropical and sub-tropical regions in the world. According to Brady et al., (2014) 96 million apparent infections per year among inhabitants of 128 countries, and nearly four billion of the population are at-risk. Dengue infection is caused by four antigenically distinct dengue virus (DENV) serotypes (DENV-1, DENV-2, DENV-3, & DENV-4) belonging to the Flaviviridae family. These viruses are transmitted through the bite of infected Aedes aegypti and Aedes albopictus female mosquitoes. Dengue incidents are being increasingly revealed in urban areas, mainly due to deficiency in water management, including improper water storage practices, and out of keeping the attention to the removal of vector breeding sites (Ramachandran et al., 2016). Since there is no antiviral treatment currently available for dengue, the precautions of dengue need control or elimination of the mosquitoes carrying out the virus that causes dengue. Therefore, adequate future management strategies must be implemented to know the dynamics of the virus, host, vector, and environmental factors especially in the context of climate variation.

In recent years, several studies have focused on the distribution of dengue. One such study (Majid et al., 2019) was performed to identify the spatial

*corresponding author: *jaminthahashini@gmail.com*, *https://orcid.org/0000-0002-4041-8458*

distribution pattern of dengue in the Seremban district in Malaysia and found that the distribution pattern is clustered. Another study (Brady et al., 2014) has been done to build a dengue transmission model to understand when and where the temperature is likely to be the significant factor that limit the rate of transmission. Their results can be used to identify the effects of temperature and other factors that may cause the expansion of dengue or its Aedes vectors. A study in Singapore by Hii et al., (2012) analysed climatological variables using the time series Poisson regression model to understand the optimal dengue forecasting period. This weather-based dengue forecasting model permits warning of 16 weeks in advance of dengue epidemics with high sensitivity and specificity.

In Sri Lanka, dengue fever was serologically confirmed in 1962 (Goto et al., 2013), and the four DENV serotypes have been co-circulating for more than 30 years (Sirisena and Noordeen, 2014). According to the data recorded in the Epidemiology Unit of the Ministry of Health in Sri Lanka, it was noted that dengue incidents were mostly reported from the Western Province, and the Colombo district was the most affected area with the highest number of reported incidents. Several studies were conducted to understand the distributional pattern of dengue incidents in Sri Lanka. A study was performed to develop and validate a forecasting model using weather variables for the Gampaha district, Sri Lanka (Withanage et al.,2018). Time-series regression analysis was used in this study, and the forecasting models allow warnings of imminent outbreaks and epidemics in advance of one month. Sun et al., (2017) have developed a spatial-temporal distribution of dengue and climatic characteristics, and two spatiotemporal clusters were detected. This study helps to understand how climatic factors impact the spatial and temporal spread of the dengue virus. The spatial and temporal distribution of Dengue in Sri Lanka was further analysed by Sirisena et al., (2017) using the Geographical Information Systems (GIS), and elucidated the association of climatological factors with dengue incidence. Wagner et al., (2020) transmission framework for the dengue virus in Sri Lanka using spatially resolved temperature and precipitation as well as the time-series Susceptible-Infected Recovered (SIR) model. Extreme weather occurrences and patterns of human mobility on the dengue outbreaks have been discussed using this SIR model.

The above studies performed so far described the mean behaviour of dengue incidents. Using such studies, we can understand and predict only the mean features of the incidence, but not the extreme cases. To get an articulate inference about the tail of the distribution, which represents the extreme events of a process, proper statistical techniques such as extreme value theory (EVT) is needed to be applied. Identifying the extreme behaviour of dengue incidents is an important step in public health surveillance and planning since peaks cause significant stress on human and public health services. Therefore, the main objective of this study was to identify the best fitted extreme distributions and to estimate future extreme events for dengue morbidity data using two approaches of EVT that is, block maxima approach and peaks over threshold approach. Identifying the association between dengue incidents and meteorological factors is also useful for the development of dengue warning systems, and hence, the administration can set out the dengue control measures promptly. A recent study (Chandrakantha, 2019) shows how climatological factors affecting the spread of dengue in the city of Colombo, Sri Lanka over the period from 2010 to 2018 using the Poisson and negative binomial regression approach. The study didn't consider the time lag and has been analysed in small geographical areas. Therefore, the second objective of this study was to understand the climatological factors on dengue incidents in seven districts of Sri Lanka; Colombo, Gampaha, Kandy, Badulla, Ratnapura, Anuradhapura, and Kurunegala using the time lag phases.

Materials and Methods

Study site and Data

Sri Lanka is a tropical country having a total land area of 64,740 km² within 5º to 10º North latitude and 79º to 82º East longitude. The country has 25 administrative districts organized into nine provinces. The monsoon winds of the Indian Ocean and the Bay of Bengal are caused for rainfall patterns in Sri Lanka. The minimum and maximum rainfall varies from 900 mm to 5000mm, and the average yearly temperature falls between the ranges from 28 to 30 °C. The island is traditionally divided into three climatic zones as dry, intermediate, and wet zone based on the seasonal rainfall. The climate in Sri Lanka can be characterized into four climatic seasons as First Inter monsoon season (March – April), Southwest monsoon season (May – September), Second Inter monsoon season (October – November), and Northeast monsoon season (December – February).

For this study, weekly dengue incidents over the period from January 2010 to December 2018 were collected from the website of the Epidemiology Unit of the Ministry of Health (MoH), Sri Lanka, and rainfall, temperature, and relative humidity data from January 2010 to December 2018 were obtained from the Meteorology Department of Sri Lanka. Rainfall amounts were measured in millimetre (mm) and the temperature was measured in Celsius (℃). Data were collected for seven administrative districts namely, Anuradhapura, Badulla, Colombo, Gampaha, Kandy, Kurunegala, and Ratnapura (Figure 1).

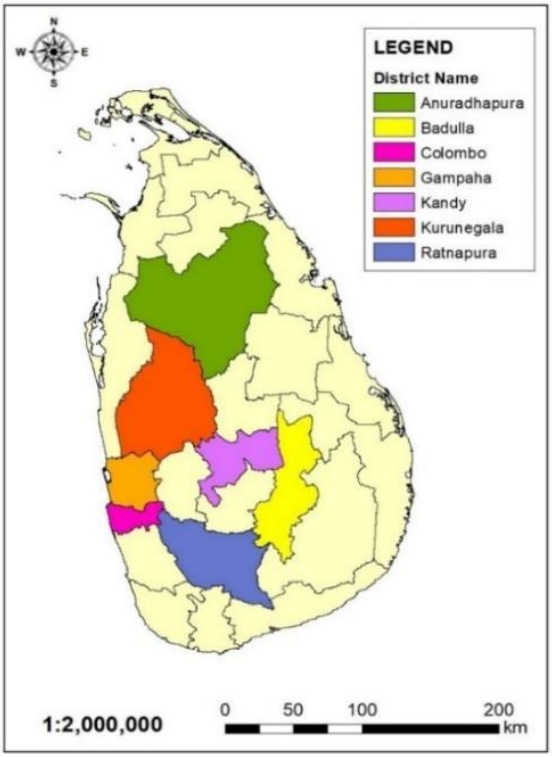


Figure 1: The map of study areas

Modelling extreme dengue incidents using Extreme Value Theory (EVT)

The EVT mainly focuses on the extreme behaviour of a process, that is the tail of the probability distribution (Coles, 2013). Two significant approaches of EVT, that is Block Maxima Method and Peaks Over Threshold (POT) method were used to identify the best-fitted extreme distributions. Then, by comparing the Root Mean Square Error (RMSE) of each fitted distribution, the best extreme distribution was selected among fitted Generalized Extreme Value (GEV) and Generalized Pareto (GP) distributions. Finally, the estimates of return levels and their 95% confidence intervals were calculated based on the selected distribution.

Block maxima method

The block maxima (BM) method splits the observation period into non-overlapping periods of equal size and extracts the maximum observation in each period. The generalized extreme value (GEV) distribution (Figure 2) was fitted using the observations extracted by the block maxima method. The cumulative distribution function of GEV has the following form:

$$F(x) = exp\left[-\left(1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\varepsilon}}\right]; if\ \ \varepsilon\ \neq 0\ ,$$

where *x*'s are the extreme values from the blocks, μ is the location parameter, σ is the scale parameter, ε is the shape parameter of the distribution, and $\left(1+\varepsilon\left(\frac{x-\mu}{\sigma}\right)\right) > 0.$

Note that the GEV distribution has three limiting distributions of extreme value depending on the value of the shape parameter (Coles, 2001):

Case 1: ε = 0 giving the light-tailed Gumbel case (EV1),

Case 2: ε > 0 giving the heavy-tailed Fréchet case (EV2)

Case 3: ε < 0 giving the short-tailed negative-Weibull case (EV3).

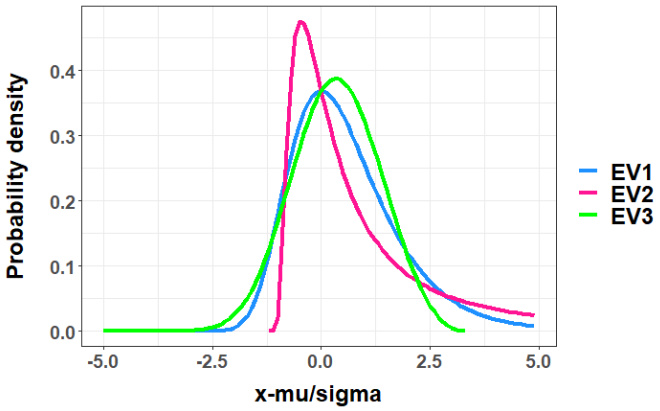


Figure 2: Probability density function of GEV distribution

Before fitting generalized extreme value distributions, it is required to test the trend, independence, and homogeneity of data of each selected block. To check these assumptions, the Mann-Kendall test (MK), Wald-Wolfowitz test (WW), and Wilcoxon test (WX) were used, respectively. The Kolmogorov-Smirnov test and Anderson-Darling test were used to test how

adequate the fitted extreme value distributions for actual data.

The Root Mean Square Error;

$(RMSE) = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\tilde{Y}_t - Y_t)^2}$ was used to select the best block size for the generalized extreme value distribution, where $\tilde{Y}_t$ is the estimated peak and Y_t is the observed peak at time point t. The block size with minimum RMSE was selected as the best block size. The GEV distribution unites the Gumbel, Fréchet, and Weibull distributions into a single-family based on its shape, location, and scale parameters to allow a continuous range of possible shapes. These three distributions are also known as type I, II, and III extreme value distributions. The GEV is equivalent to type I, II, and III, respectively, when a shape parameter is equal to 0, greater than 0, or lower than 0. Therefore, after selecting the best-fitted model, the type of the GEV can be identified based on its parameters.

Peaks over threshold method

In the Peaks Over Threshold (POT) method, the data were collected over some specific high threshold value. The issue of how to select the threshold was similar to that of selecting the block size in block maxima method that both imply a balance between bias and variance. A lower threshold level leads to failure in the asymptotic approximation of the model and a high threshold level provides few observations with high variance. The choice of appropriate thresholds in this study was based on the mean residual plot. Threshold values were selected for every region, and the possible threshold was the point when the mean residual plot shows linearity, and estimated parameters look stable at different thresholds (Bommier, 2014). For several possible threshold values, the empirical mean residual function was fitted to select the best threshold value. Then, the Generalized Pareto (GP) distribution (Figure 3) was fitted using the observations which were greater than the selected threshold value. The cumulative distribution function of GP distribution is defined below:

$F(x, \varepsilon, \sigma, \mu) = 1 - \left[1 + \varepsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\varepsilon}}$; $if\ \ \varepsilon \neq 0$,

where $X > U$ is a random variable with a given threshold U, μ is the location parameter, $\sigma > 0$ is the scale parameter, ε is the shape parameter, and $1 + \varepsilon\left(\frac{x-u}{\sigma}\right) > 0$.

Based on the fitted GP distribution, the stability plots were obtained for the scale and shape parameters to select the best threshold value. In the POT method, the exceedances should be mutually independent (Omey et al., 2009). To obtain a set of threshold excesses that are approximately independent and be able to apply the Peaks over threshold method, the declustering method was used to filter the dependent observations from whole exceedances. The underlined assumptions of the GP distribution corresponding to the selected threshold are the trend, independence, and homogeneity of data. Therefore, to check these assumptions, the Mann-Kendall test (MK), Wald-Wolfowitz test (WW), and Wilcoxon test (WX) were used, respectively. The Kolmogorov-Smirnov test and Anderson-Darling test were used to test how adequate the fitted GP distribution for actual data.

The GP distribution has three basic forms based on its shape parameter, and each corresponds to a limiting distribution of exceedance data from a different class of distributions.

Case 1: The shape parameter $\varepsilon \rightarrow 0$ (GP1)

In this case, the tails of GP distribution decreases exponentially, and the GP distribution follows an exponential distribution with mean σ. Then, $F(x, \sigma, \mu) = 1 - exp\left(-\frac{x-\mu}{\sigma}\right)$.

Case 2: The shape parameter $\varepsilon > 0$ (GP2)

When $\varepsilon > 0$, tails of GP distribution decreases as a polynomial, and the GP distribution follows the Pareto distribution.

Case 3: The shape parameter $\varepsilon < 0$ (GP3)

When $\varepsilon < 0$, tails of GP distribution have finite upper end points, and the GP distribution follows the Beta distribution.

Therefore, the relevant form of the GP distribution can be identified based on the shape parameter of the best-fitted GP distribution.

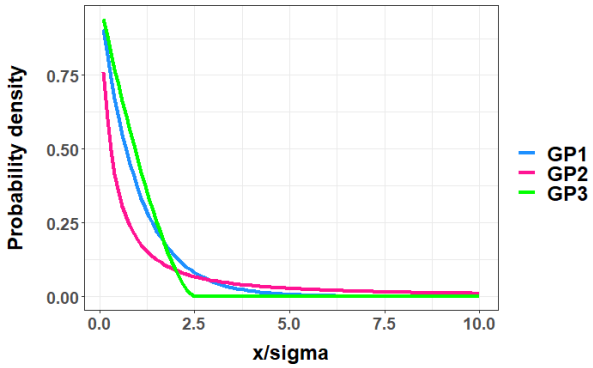


Figure 3: Probability density function of GP distribution

Return levels for GEV and GP distributions

One of the most important ultimate outcomes in data analysis related to extreme value theory is the future prediction of extreme quantiles and tail probabilities.

Extreme quantiles that occur with a certain return level can be estimated using the fitted extreme distributions. The return value is defined as a value that is expected to be equalled or exceeded on average once every interval of time (T) with probability $p = \frac{1}{T}$.

For GEV distribution, the return level (Z_p) with return period $T = \frac{1}{p}$ is obtained by,

$$Z_p = \{\mu - \frac{\sigma}{\varepsilon}\left[1 - \left(-log(1-p)\right)^{\varepsilon}\right] \; ; if \; \varepsilon \neq 0$$

where μ is the location parameter, σ is the scale parameter, and ε is the shape parameter of the GEV distribution.

For the GP distribution, the return level (X_m),

$X_m = \{u + \frac{\sigma}{\varepsilon}[(\zeta u) - 1] \;\; ; \;\; if \;\; \varepsilon \neq 0$

is described as the extreme level that is exceeded on average once every m observations, where u represents the threshold value, ζu is the probability of the occurrence of exceedance of threshold u, σ is the scale parameter, and ε is the shape parameter.

Fitting a weather-based dengue forecasting model using Generalized Linear Negative Binomial (GLNB) distribution

In order to identify the association between climatological factors and dengue incidents, empirical models can be developed. The outcome variable; dengue incidents are measured as counts for which the distribution is over-dispersed. Thus, the Generalized Linear Negative Binomial (GLNB) model was used to fit the empirical models. Under the influence of climatic factors, it takes seven to 45 days for an adult mosquito to develop from an egg (Ramachandran et al., 2016). Hence, the influence of climatic factors was expected to manifest with a lag of one to three months. Four models were developed using average monthly rainfall, temperature, and humidity as the independent variables and log of monthly dengue incidents as the dependent variable. The first model was developed without time lag and the other three models were developed by considering the time lag of 1, 2, and 3 months, respectively. One-way analysis of variance was used to test whether each of the climate variables differs significantly between monsoon seasons. Since a significant correlation was observed in the variability of dengue incidents across monsoon seasons, an additional categorical variable "monsoon season" was added and conducted the same analysis. The above models were fitted separately for Anuradhapura, Badulla, Colombo, Gampaha, Kandy, Kurunegala, and Ratnapura districts. The best-fitted distribution was identified by comparing the Akaike information criterion (AIC) and Bayesian information criterion (BIC), and the accuracy of the models for predicting outbreaks was assessed through the receiver operating characteristic (ROC) curve.

3. Results and Discussion

Descriptive Statistics of the variables

As a preliminary analysis, the variation of dengue incidents, rainfall, temperature, and humidity among inter-monsoon seasons for the selected districts are considered. Since the mean and standard deviations are different among inter-monsoon seasons, coefficient of variation was taken as a measure to compare the variation. In Appendix, Table A1, coefficients of variation of dengue incidents are shown for each inter-monsoon season for the selected districts. For each district, the highest variation is shown in bold. According to the results, Colombo, Kandy, Badulla and Ratnapura districts show the highest variability of dengue incidents in the Southwest monsoon season. Further, Gampaha and Kurunegala districts have the highest variability of dengue incidents in the Second- Inter monsoon season, and Anuradhapura district show the highest variability of dengue incidents in the Northeast monsoon season.

Further, Table A2 in the appendix presents the coefficient of variation of rainfall, temperature, and humidity for each inter-monsoon season for the selected districts. The highest variation for each case is shown in bold. According to the results in table A2, variation in rainfall is higher than that of temperature and humidity. In the Northeast monsoon season Colombo, Gampaha, Kandy, Kurunegala, Ratnapura, and in the Southwest monsoon season Badulla and Anuradhapura show the highest variability in rainfall.

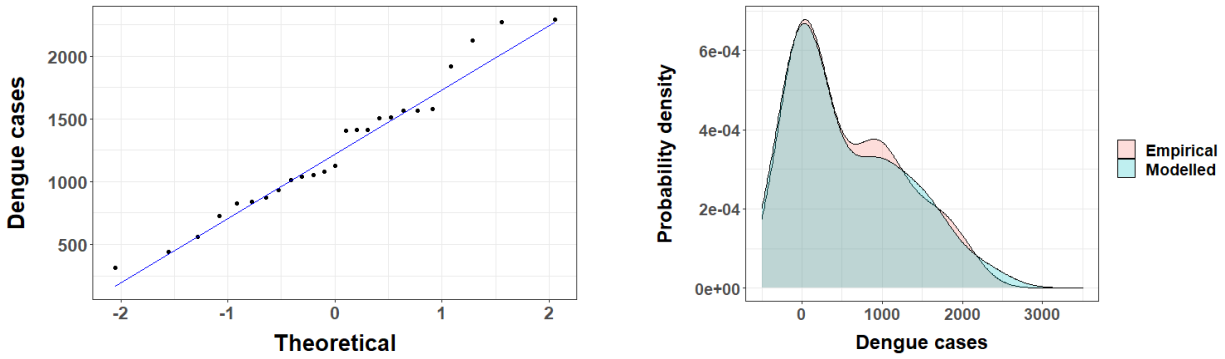
Fitting a GEV distribution for dengue incidents

Four block sizes (4, 8, 12, & 16) were selected to fit a GEV. Before fitting the GEV distribution, the assumptions required to fit a GEV were tested. According to the results, it was noted that Mann-Kendall, Wald-Wolfowitz, and Wilcoxon tests are not significant at 5% significance level for block sizes of 12 and 16. Hence, block sizes of 12 and 16 were used for model fitting since the assumptions required to fit a GEV distribution are valid. Then, to

test how adequate the fitted GEV distributions for block sizes of 12 and 16, Anderson-Darling (AD) test and Kolmogorov- Smirnov (KS) tests were performed and obtained the RMSE values. Table 1 shows the *p*-values for AD and KS tests, the estimated parameters, their confidence intervals, and the respective RMSE values for the fitted GEV distributions.

According to the results, both block sizes, 12 and 16, have a better fit for GEV distribution as Anderson-Darling and Kolmogorov- Smirnov tests are not significant at 5% significance level. However, when comparing the RMSE values, block size 16 has the minimum.

Therefore, block size 16 was selected as the suitable block size, and the GEV distribution with relevant parameters can be used to identify the extreme behaviour of dengue incidents in the study areas. The estimated shape parameter of the GEV is negative (-0.1944), while the confidence interval can be zero as well (Table 1). Therefore, the GEV is equivalent to the light-tailed Gumbel case (EV1). Quantile and density plots obtained from dengue data with block size 16 are shown in Figure 4.



(a) (b)

Figure 4: Quantile (a) and density (b) plots for GEV distribution for block size 16

Table 1: Estimated parameters, confidence intervals, adequacy tests, and RMSE for fitted GEV distributions

GEV		Estimated Parameters			Adequacy (*p*-value)		RMSE
Block	N	shape	scale	location	KS	AD	GEV
12	33	-0.004 (-0.331, 0.323)	417.363 (292.313, 542.412)	868.409 (701.399, 1035.418)	0.951	0.882	7.898
16	25	-0.194 (-0.526, 0.137)	488.535 (330.348, 646.723)	1048.983 (830.137, 1267.829)	0.854	0.658	6.883

Figure. 4(a) indicates the comparison of the model and the empirical data quantiles. A quantile plot that deviates considerably from a straight line suggests that the model assumptions may be invalid for the plotted data. However, in Figure 4(a), only few points with deviations can be identified in the quantile plot. Figure 4(b) shows the behaviour of empirical and fitted model for GEV distribution, which shows that the two densities are approximately close. Hence, the data approximately well-fitted for the GEV distribution with block size 16.

Fitting a GP distribution for dengue incidents

Before fitting a GP distribution, the threshold values were selected using the mean residual life plot. Figure. 5 shows the mean residual life plot with 95% confidence interval.

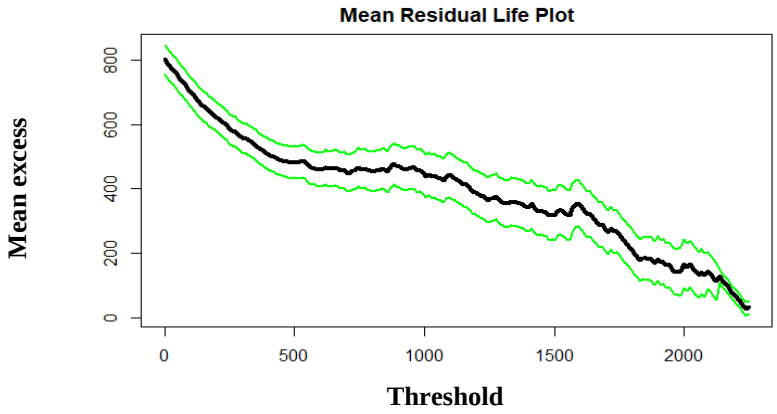
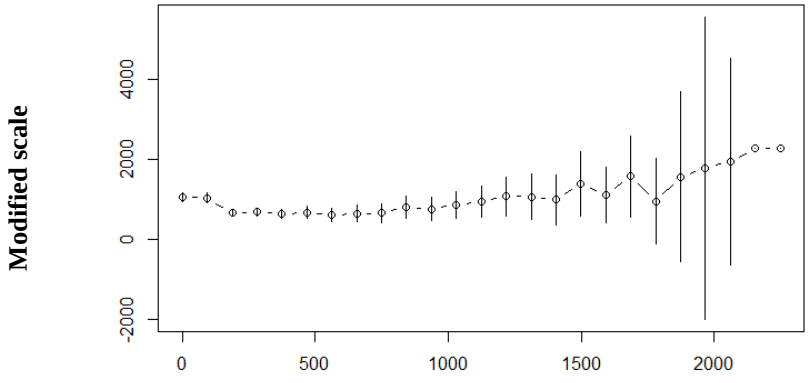
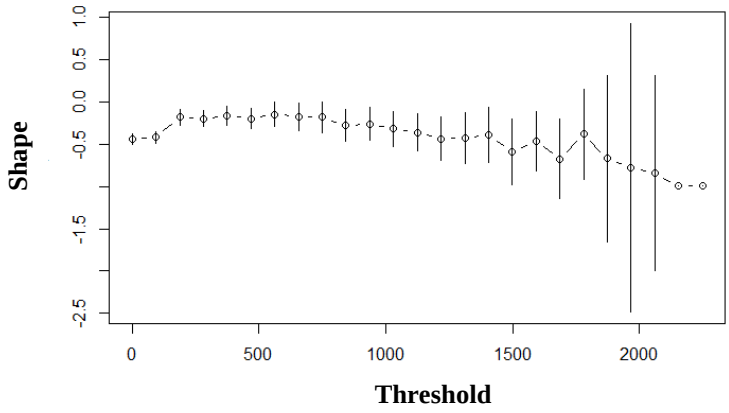


Figure 5: Mean Residual life plot of dengue cases

According to Fig. 5, a linearity can be observed in between 500 and 1000. Then, to identify a suitable cut-off value, the following stability plots for the scale and shape parameters of the fitted GP distribution were used.



(a) Stability plot of scale parameter



(b) Stability plot of shape parameter

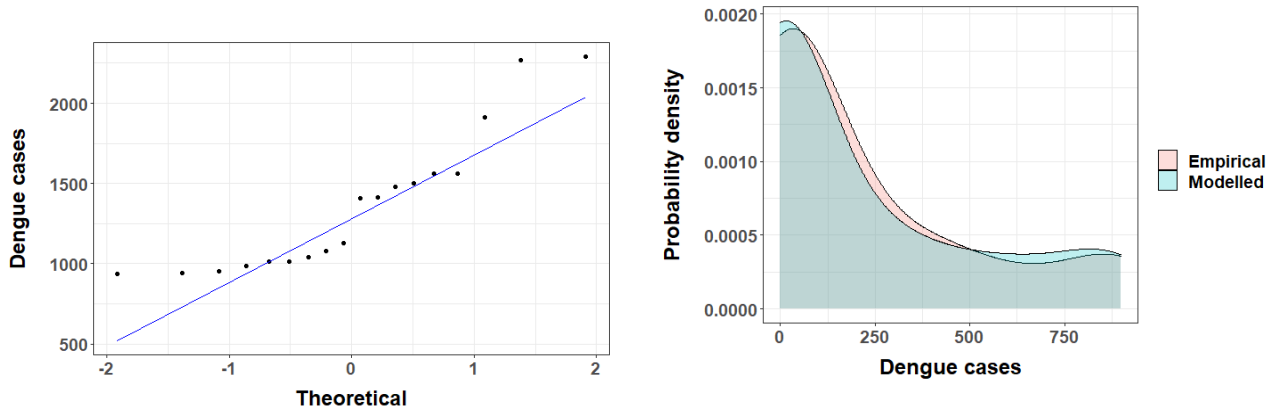
Figure 6: Stability plot for fitted GP model

According to the above stability plots, the linearity breaks down at a value near 900. Therefore, 900 was selected as the best threshold value to fit the GP distribution. Since the data points above the threshold value 900 are clustered, the declustering method was used to obtain a set of threshold excesses that are approximately independent. The estimated parameters for declustered dengue exceedances, their confidence intervals, and the *p*-values obtained for testing adequacy (Anderson Darling test and the Cramer-Von Misses test) for the fitted distribution are shown in the Table 2.

Table 2: Estimated parameters, confidence intervals, adequacy tests, and RMSE for fitted GP distribution

GPD		Estimated Parameters		Adequacy (*p*-value)		RMSE
Threshold	n	Shape	Scale	AD	Cramer-von Misses	
900	18	-0.190	553.432	0.214	0.223	8.747
		(-0.855, 0.474)	(108.094, 998.769)			

The *p*-values of the Anderson-Darling (AD) and Cramer-Von Misses tests are not significant at 5% significance level (Table 2), which indicates that the data follows a GP distribution. Although the estimated shape parameter is -0.190, the confidence interval suggests that it can even be zero. Therefore, the fitted distribution is an exponential type distribution.



(a) (b)

Figure 7: Quantile (a) and density (b) plots for GP distribution

Fig. 7 shows the quantile plot and density plot for the fitted GP distribution for threshold 900. Since most of the scatter points close to the reference line of the quantile plot, and the empirical and modelled densities are approximately the same as shown in Fig.7(b). Thus, the fitted GP distribution is a better fit for the weekly dengue incidents in the study area.

Predicting extreme dengue incidents

To select the most appropriate distribution among the two fitted distributions GEV and GP, their RMSE values can be compared. According to Tables 1 and 2, the GEV distribution has the minimum RMSE (6.883), and hence, the GEV distribution can be selected as the best-fitted distribution to estimate extreme dengue incidents in the study area. Now, using this distribution, the return level plot, return level estimates, and their 95% confidence intervals at different return periods can be obtained.

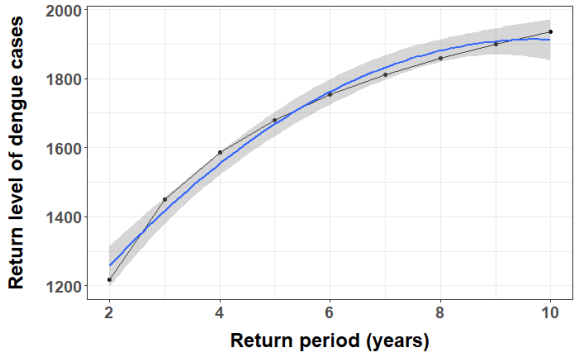


Figure 8: Return level plot

Figure 8 shows the return level plot, which includes the estimated return levels and a 95% confidence band for several years. Since there is no study to predict future extreme dengue quantiles in Sri Lanka, these results would lead to considerable potential in aiding health planners in the risk management of dengue.

Table 3: GEV distribution return level estimates and confidence intervals

Return Period	2 years	3 years	4 years	5 years	6 years	7 years
Probability	0.50	0.33	0.25	0.20	0.17	0.14
Lower CI	986.874	1203.802	1331.951	1419.560	1483.749	1532.839
Return level	1217.493	1448.923	1585.108	1680.340	1752.857	1811.013
Upper CI	1448.111	1694.044	1838.266	1941.119	2021.965	2089.187

Note that all estimated return levels are within the 95% confidence band, and the estimated return levels increase with the increase of the return period. The Table 3 shows the estimated return levels, and their 95% confidence intervals at different return periods for 16 weekly maxima.

According to the estimated return levels, a maximum of 1217 dengue incidents are expected to appear once in every 2 years, over a long period. It can vary between 986 and 1448. Similarly, a maximum of 1448, 1585, 1680, 1752, 1811 dengue incidents are expected to appear once in every 3, 4, 5, 6, and 7 years (Figure 8).

Fitting a weather-based model using Generalized Linear Negative Binomial (GLNB)

To fit GLNB models, log of dengue counts is taken as the response variable, and rainfall, temperature, and humidity as predictor variables. One-way analysis of variance (ANOVA) was used to test whether each of the climate variables (rainfall, temperature, humidity) differs significantly among the four seasons ((First Inter Monsoon, South-West Monsoon, Second Inter Monsoon, North-East Monsoon). The following table shows the results of one-way ANOVA.

Table 4: ANOVA results for climatic factors on monsoon seasons.

District	ANOVA results (p-value)		
	Rainfall	**Temperature**	**Humidity**
Colombo	$3.21e^{-01}$	$2.11e^{-10}$ ***	$7.78e^{-4}$ ***
Gampaha	$4.36e^{-01}$	$7.78e^{-13}$ ***	$1.51e^{-3}$**
Kandy	$4.20e^{-02}$ *	$1.33e^{-11}$ ***	$3.60e^{-2}$*
Badulla	e^{-03} **	$4.64e^{-16}$ ***	$1.73e^{-09}$ ***
Kurunegala	$3.81e^{-01}$	$9.86e^{-16}$ ***	$8.91e^{-1}$
Ratnapura	e^{-02} *	$8.25e^{-07}$ ***	$4.83e^{-1}$
Anuradhapura	$7.50e^{-01}$	$2.00e^{-16}$ ***	$2.38e^{-10}$ ***

* $p \leq 0.05$, ** $p \leq 0.01$, and, *** $p \leq 0.001$

The results in Table 4 indicate that the climatic variables are significantly different between seasons except for the rainfall of Colombo, Gampaha, Kurunegala & Anuradhapura, and for the humidity of Kurunegala and Ratnapura.

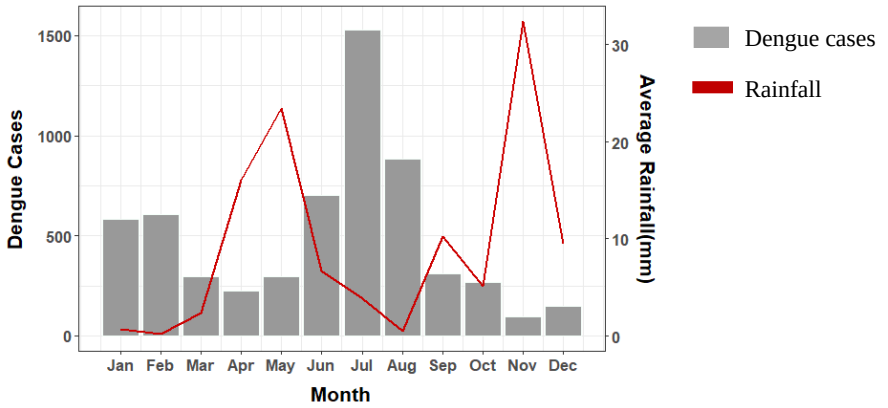


Figure 9: Monthly average of dengue cases, and monthly rainfall in Colombo from 2010 to 2017

The months with higher rainfall do not have a higher number of dengue incidents (Figure 9), while after one, two, or three months lag of the rainy months, there is an increment in dengue incidents (Figure 9). This is because, under the influence of climatic factors, it takes 7 to 45 days for an adult mosquito to develop from an egg. By considering these circumstances, empirical models were fitted by adding a lag (0, 1, 2, & 3) as a variable in the GLNB model. Two types of GLNB models were fitted separately, to understand how the climatic factors affect seasonally on dengue incidences. First, the GLNB models were fitted without considering the seasonality effect of the climatic variables. Then, the GLNB models were again fitted by adding the monsoon seasons (First Inter Monsoon, South-West Monsoon, Second Inter Monsoon, North-East Monsoon) as covariates, and compared the two types of models for prediction accuracy. For each district, four models were fitted by considering time lag (0, 1, 2, & 3), and the best-fitted model was selected using AIC and BIC values. The Table 5 shows the best-fitted empirical models without considering the seasonality effect of the climatic variables, and the significant coefficient of each variable is shown in bold.

Table 5: Summary of fitted empirical models without monsoon seasons

District	Lag (month)	Variables			
		Constant	**Rainfall**	**Temperature**	**Humidity**
Colombo	3	1.537 ($4.81e^{-01}$)	**0.034** **($4.00e^{-03}$ **)**	**0.231** **(e^{-03} **)**	-0.019 ($3.01e^{-01}$)
Gampaha	0	**7.255** **($5.00e^{-03}$**)**	**-0.045** **($5.00e^{-03}$ **)**	-0.102 ($2.52e^{-01}$)	0.027 ($9.90e^{-02}$)
Kandy	2	**-7.288** **($3.00e^{-03}$**)**	0.008 ($6.83e^{-01}$)	**0.353** **($1.45e^{-07}$ ***)**	**0.047** **($7.56e^{-03}$ **)**
Badulla	2	**-11.150** **($1.50e^{-04}$ ***)**	-0.008 ($7.39e^{-01}$)	**0.491** **($2.30e^{-09}$ ***)**	**0.053** **($3.07e^{-03}$ **)**
Kurunegala	2	-2.169 ($3.74e^{-01}$)	0.020 ($2.09e^{-01}$)	**0.162** **($2.24e^{-02}$ *)**	**0.039** **($6.75e^{-03}$ **)**
Ratnapura	3	**-8.039** **($4.67e^{-02}$ *)**	0.001 ($9.72e^{-01}$)	**0.516** **($8.86e^{-06}$ ***)**	-0.013 ($5.17e^{-01}$)
Anuradhapura	2	-1.887 ($4.51e^{-01}$)	0.001 ($9.86e^{-01}$)	0.066 ($2.81e^{-01}$)	**0.056** **($2.29e^{-05}$ ***)**

* $p \leq 0.05$, ** $p \leq 0.01$, and, *** $p \leq 0.001$

The results in Table 5 indicate that Anuradhapura, Badulla, Kurunegala, and Kandy districts have the best model at lag 2 whereas Ratnapura and Colombo districts have lag 3, and for Gampaha district, it is lag 0. Similarly, four models were fitted for each district by adding monsoon season and time lag as covariates, and the best model was selected based on AIC and BIC values. The results of the best-fitted GLNB models are given in the table 6. The results in Table 6 indicate that Anuradhapura, Badulla, and Kandy districts show their best model at lag 2 whereas Colombo and Gampaha show lag 0 models and Kurunegala shows lag 1 model. The best empirical model at lag 3 is for the Ratnapura district.

Retrospective validation of the empirical models was done with 2018 data, and the accuracy of the empirical models for predicting outbreaks was assessed through the receiver operating characteristic (ROC) curve (Table 7).

Table 6: Summary of fitted empirical models with monsoon seasons

District	Lag (month)	Variables				
		Constant	**Rainfall**	**Temperature**	**Humidity**	**Season**
Colombo	0	-1.384 ($6.19e^{-01}$)	**-0.059** **($2.01e^{-08}$ ***)**	0.130 ($1.15e^{-01}$)	**0.057** **($6.30e^{-04}$ ***)**	**0.256** **($3.19e^{-04}$ ***)**
Gampaha	0	-3.594 ($3.56e^{-01}$)	**-0.046** **($2.23e^{-03}$ **)**	0.190 ($1.09e^{-01}$)	**0.052** **($1.62e^{-03}$ **)**	**0.327** **($2.54e^{-04}$ ***)**
Kandy	2	**-7.287** (**$1.55e^{-02}$ ***)	0.008 ($6.87e^{-01}$)	**0.353** (**$5.17e^{-05}$ *****)	**0.048** (**$8.19e^{-03}$ ****)	-0.001 ($9.10e^{-01}$)
Badulla	2	**-12.811** **($7.65e^{-05}$ ***)**	-0.009 ($6.93e^{-01}$)	**0.551** **($1.47e^{-08}$ ***)**	**0.053** **($3.18e^{-03}$ **)**	0.121 ($2.39e^{-01}$)
Kurunegala	1	**-10.033** **($5.89e^{-03}$ **)**	-0.009 ($5.42e^{-01}$)	**0.347** **($1.16e^{-03}$ **)**	**0.066** **($3.54e^{-06}$ ***)**	**0.398** **($1.68e^{-04}$ ***)**
Ratnapura	3	-8.038 ($1.07e^{-01}$)	0.001 ($9.73e^{-01}$)	**0.516** **($3.75e^{-04}$ ***)**	-0.013 ($5.22e^{-01}$)	-0.001 ($9.10e^{-01}$)
Anuradhapura	2	-4.998 ($1.38e^{-01}$)	0.002 ($8.48e^{-01}$)	0.159 ($8.34e^{-02}$)	**0.057** **($1.11e^{-05}$ ***)**	0.141 ($1.96e^{-01}$)

* $p \leq 0.05$, ** $p \leq 0.01$, and, *** $p \leq 0.001$

According to the results in Table 7, the empirical model with the higher accuracy to predict the relevant dengue outbreak for each district is shown in bold. For example, the empirical model with monsoon season for the Colombo district shows approximately 81% accuracy to predict dengue outbreaks (greater than 930 incidents). To understand the significant variables, the effect of the four monsoon seasons, and the models with the highest prediction accuracy. Further, the results in Table 5-7 are summarized in Table 8.

In all models, the monsoon season is not significant except for the models of Colombo, Gampaha, and Kurunegala districts for which the covariate monsoon season has a significant positive effect. Further, the models of Colombo and Gampaha have a significant negative effect on rainfall. For example, the coefficient of the rainfall for GLNB model with the highest prediction accuracy of Colombo is -0.05942. Since log ($dengue\ count$) is the response variable, then one unit of decrease of rainfall, the dengue counts will increase by 34.7%. Note that, Colombo and Gampaha districts are in the wet zone of Sri Lanka. The humidity shows a significant positive effect for all models with high accuracy, except for Ratnapura, which is located partly in the wet climatic zone and partly in the intermediate zone. The rainfall in Ratnapura is around 4460 mm per year, with precipitation even during the driest month. For Ratnapura, the only significant positive effect is

temperature. Further, the humidity is the only significant positive effect in GLNB model for the Anuradhapura district which is in the dry zone. Both humidity and temperature are the positive significant effects for models of Kandy, Badulla, and Kurunegala, where Badulla and Kurunegala districts are in the intermediate zone, and Kandy mostly belongs to the wet zone but partly to the intermediate zone.

Table 7: Receiver operating characteristic (ROC) curve results

District	Outbreak	AUC (Area under the curve)	
		Without monsoon seasons	With monsoon seasons
Colombo	930	0.438	**0.813**
Gampaha	400	0.567	**0.733**
Kandy	150	**0.792**	**0.792**
Badulla	100	**0.667**	0.583
Kurunegala	200	0.667	**0.769**
Ratnapura	350	**0.667**	**0.667**
Anuradhapura	45	0.700	**0.733**

The models with the highest prediction accuracy are shown in bold.

Table 8: Summary of results for GLNB models

	Model without monsoon	Model with monsoon season
District	**Lag (month) and significant variables**	**Lag (month) and significant variables**
Colombo	3 (Rainfall, Temperature)	**0** **(-Rainfall, Humidity, Season)**
Gampaha	0 (Constant, -Rainfall)	**0** **(-Rainfall, Humidity, Season)**
Kandy	**2** **(-Constant, Temperature, Humidity)**	**2** **(-Constant, Temperature, Humidity)**
Badulla	**2** **(-Constant, Temperature, Humidity)**	2 (-Constant, Temperature, Humidity)
Kurunegala	2 (Temperature, Humidity)	**1** **(-Constant, Temperature, Humidity, Season)**
Ratnapura	**3** **(-Constant, Temperature)**	**3** **(Temperature)**
Anuradhapura	2 (Humidity)	**2** **(Humidity)**

Models with highest prediction accuracy are shown in bold

Conclusions

This study shows that the GEV model fits well than the GP distribution for the extreme dengue incidents in Sri Lanka. Those models can act as an early warning system for enhancing measures of dengue control to reduce the size of an outbreak. Furthermore, the study found that climatological factors can serve as important components for generating an uncomplicated and timely dengue forecasting system. When considering GLNB models with monsoon season as a covariate, lag 0 model is the suitable model for Colombo and Gampaha districts with the highest prediction accuracy.

Similarly, lag 1 model is the suitable for Kurunegala whereas lag 2 model is the best for Anuradhapura having the highest accuracy when monsoon season is added as a covariate. For Badulla district, lag 2 model without having monsoon season as a covariate shows highest prediction accuracy. Note that the prediction accuracy is the same for the models with or without having the monsoon season as a covariate for Kandy and Ratnapura districts. The significant climatic factors are different from district to district, and the humidity and/or temperature is significant for most districts. The development of a weather-based forecasting model can assist the prevention and surveillance of dengue incidents in several ways. It decreases disease transmission and possibly the resulting mortality, leading to a reduction in the health care burden and operating costs.

Acknowledgement

We would be grateful to the Meteorology Department of Sri Lanka for providing climatology data.

References

Bommier, E. (2014). Peaks-Over-Threshold Modelling of Environmental Data. Department of Mathematics Uppsala University, U.U.D.M. P(September), 35.

Brady, O. J., Golding, N., Pigott, D. M., Kraemer, M. U. G., Messina, J. P., Reiner, R. C., Scott, T. W., Smith, D. L., Gething, P. W., & Hay, S. I. (2014). Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus transmission. Parasites and Vectors, 7(1), 1–17.

Chandrakantha, L. (2019). Statistical Analysis of Climate Factors Influencing Dengue Incidences in Colombo, Sri Lanka: Poisson and Negative Binomial Regression Approach. International Journal of Scientific and Research Publications (IJSRP), 9(2), p8616.

Coles, S. (2013). An Introduction to Statistical Modeling of Extreme Values. 1st ed. London: Springer London, Limited, pp.45-72.

Goto, K., Kumarendran, B., Mettananda, S., Gunasekara, D., Fujii, Y., & Kaneko, S. (2013). Analysis of Effects of Meteorological Factors on Dengue Incidence in Sri Lanka Using Time Series Data. PLoS ONE, 8(5), 1-8.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., & Rocklöv, J. (2012). Forecast of Dengue Incidence Using Temperature and Rainfall. PLoS Neglected Tropical Diseases, 6(11), 1-9.

Majid, N. A., Nazi, N. M., & Mohamed, A. F. (2019). Distribution and spatial pattern analysis on dengue incidents in Seremban District, Negeri Sembilan, Malaysia. Sustainability (Switzerland), 11(13), 1-14.

Omey, E., Mallor, F., & Eulalia Nualart. (2009). An introduction to statistical modelling of extreme values. Application to calculate extreme wind speeds.

Ramachandran, V. G., Roy, P., Das, S., Mogha, N. S., & Bansal, A. K. (2016). Empirical model for estimating dengue incidence using temperature, rainfall, and relative humidity: a 19-year retrospective analysis in East Delhi. Epidemiology and Health, 38, 1-8.

Sirisena, P. D. N. N., & Noordeen, F. (2014). Evolution of dengue in Sri Lanka-changes in the virus, vector, and climate. International Journal of Infectious Diseases, 19(1), 6–12.

Sirisena, P., Noordeen, F., Kurukulasuriya, H., Romesh, T. A., & Fernando, L. K. (2017). 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, 12(1), 1-14.

Sun, W., Xue, L., & Xie, X. (2017). Spatial-temporal distribution of dengue and climate

characteristics for two clusters in Sri Lanka from 2012 to 2016. Scientific Reports, 7(1), 1–12

Wagner, C. E., Hooshyar, M., Baker, R. E., Yang, W., Arinaminpathy, N., Vecchi, G., Metcalf, C. J. E., Porporato, A., & Grenfell, B. T. (2020). Climatological, virological and sociological drivers of current and projected dengue fever outbreak dynamics in Sri Lanka. Journal of the Royal Society, Interface, 17(167), 1-15.

Withanage, G. P., Viswakula, S. D., Nilmini Silva Gunawardena, Y. I., & Hapugoda, M. D. (2018). A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. Parasites and Vectors, 11(1), 1–10.

Appendix

Table A1: Coefficient of variation of dengue incidents for each inter-monsoon season for the selected districts

District	Inter-monsoon season			
	FI	SW	SI	NE
Colombo	43.34	**57.1**	50.95	55.13
Gampaha	31.34	62.54	**74.46**	50.94
Kandy	38.44	**93.11**	62.76	57.38
Badulla	45.63	**95.63**	83.81	84.87
Kurunegala	59.44	66.04	**90.07**	81.98
Ratnapura	53.34	**76.08**	51.28	52.72
A'pura	74.25	65.24	38.95	**85.51**

FI= First Inter, SW= Southwest, SI= Second Inter, NE= Northeast

Table A2: Coefficient of variation of rainfall, temperature and humidity for each inter-monsoon season for the selected districts

District	Rainfall				Temperature				Humidity			
	FI	SW	SI	NE	FI	SW	SI	NE	FI	SW	SI	NE
Colombo	64.95	93.63	64.02	**95.17**	2.13	1.69	1.37	**3.77**	3.75	2.53	2.55	**5.15**
Gampaha	71.34	98.56	53.24	**114.34**	2.11	1.69	1.57	**2.28**	4.89	2.74	3.38	**7.04**
Kandy	60.80	76.35	9.86	**100.52**	2.43	1.73	**6.18**	3.04	8.10	4.19	6.16	**8.67**
Badulla	70.99	**86.26**	46.58	81.40	**3.88**	1.55	2.86	2.61	5.84	7.94	**10.14**	5.5
Kurunegala	71.12	98.82	47.77	**127.4**	2.61	2.06	2.80	**3.54**	9.95	5.34	6.31	**10.9**
Ratnapura	51.33	49.55	43.90	**61.65**	2.20	1.67	1.28	**2.49**	6.44	3.93	3.74	**9.66**
A'pura	77.35	**154.86**	25.28	110.66	2.61	1.46	**3.16**	2.57	6.23	6.66	**10.90**	7.66