Received: 13 July 2020 | Accepted: 7 May 2021

DOI: 10.1111/rssa.12714

ROYAL STATISTICAL SOCIETY
Series A
Statistics in Society

ORIGINAL ARTICLE

# An ensemble method for early prediction of dengue outbreak

**Soudeep Deb**[1] | **Sougata Deb**[2]

[1]Indian Institute of Management Bangalore, Bannerghatta Main Road, Billekahalli, Bangalore, Karnataka, India

[2]Institute of Systems Science, National University of Singapore, 29 Heng Mui Keng Terrace, Singapore, Singapore

**Correspondence**
Soudeep Deb, Indian Institute of Management Bangalore, Bannerghatta Main Road, Billekahalli, Bangalore, Karnataka, India.
Email: soudeep@iimb.ac.in

**Abstract**

Predicting a dengue outbreak well ahead of time is of immense importance to healthcare personnel. In this study, an ensemble method based on three different types of models has been developed. The proposed approach combines negative binomial regression, autoregressive integrated moving average model and generalized linear autoregressive moving average model through a vector autoregressive structure. Lagged values of terrain and climate covariates are used as regressors. Real-life application using data from San Juan and Iquitos shows that the proposed method usually incurs a mean absolute error of less than 10 cases when the predictions are made 8 weeks in advance. Furthermore, using model confidence set procedure, it is also shown that the proposed method always outperforms other candidate models in providing early prediction for a dengue epidemic.

**KEYWORDS**

epidemic, ensemble forecasting, infectious disease, model confidence set, time series



## 1 | INTRODUCTION

Dengue is a major public health concern in most tropical and sub-tropical countries. It is a vector-borne infection which causes flu-like illness, fever and severe pain in the body. In addition, dengue haemorrhagic fever (DHF) is extremely complicated, which can cause haemorrhage and internal organ failures (Ahmed et al., 2020). Such complications often require hospitalization and focused medical care for dengue and DHF patients. Currently, treatment is generally supportive while vaccines continue to be under development (Buczak et al., 2018). While it is endemic to more than 100 countries,

these areas periodically experience even higher risk when vector population increases in proximity to human habitation. Environmental factors, such as hot and humid climate, moderate rainfall etc. are known to influence the prevalence of mosquito vectors (Wu et al., 2007). In such cases, the vector population generally undergoes a breeding and maturity period of 3 to 6 weeks before they start infecting humans actively, cf. Ebi and Nealon (2016). In another relevant study, Yang et al. (2009) showed that temperature has a significant effect on the life cycle of *Aedes aegypti*, the main vector behind dengue. Similar results were discussed in Kearney et al. (2009) and Estallo et al. (2012) as well.

With the dengue outbreaks becoming increasingly common across the world and subsequently leading to unrelenting burden on the health infrastructure and services (Bowman et al., 2016), understandably there have been a large number of studies to predict dengue occurrences and outbreaks using data from various parts of the world. However, most of these research findings are either too specific or the accuracy of results, based on empirical validation, is not strong enough. This is particularly evident for outbreaks and epidemics, generally defined by a prolonged multi-week period with an elevated count of patients than other periods. While the majority of the researches can model and forecast the regular periods accurately (e.g. Bowman et al., 2016; Buczak et al., 2018; Guo et al. 2017; Kilicman 2018; Wu et al., 2007), they often fail to obtain similar performance levels for the outbreak periods (Buczak et al., 2018; Kilicman 2018). In this paper, we attempt to bridge that gap by focusing on developing a robust and superior forecasting methodology, that relies on an ensemble technique of three separate models. The method performs consistently well across various types of periods and different locations. Iquitos (Peru) and San Juan (Puerto Rico) have been chosen for application and validation of the proposed methodology, since rich data on both locations are available from public sources (DengAI, 2017).

Remaining sections in this paper are structured as below. Section 2 provides a detailed review of relevant literature to understand their strengths, weaknesses and to finally combine these learnings into the proposed methodology. In Section 3, after defining some notations, three different models, the ensembling strategy and the implementation details are discussed. Section 4 explains the data and the key results based on the data from Iquitos and San Juan, along with suitable benchmark results. Finally, Section 5 discusses the findings, draws key conclusions and briefly summarizes the scope of future research on this topic.

## 2 | LITERATURE REVIEW

Studies related to dengue outbreak modelling largely follow one of the three methodologies *viz*. time series forecasting, epidemiological modelling or regression-based predictions. In recent times, various ensembles combining one or more of the above types have also been proposed.

Time series approaches, either using covariates or by using the past and recent counts of dengue cases alone, are generally found to be the most promising among all the methodologies. Such time series models are often found to supersede other model-based approaches, or have been key components in ensemble approaches using a collection of different forecasting techniques. Luz et al. (2008), Dom et al. (2013), Chakraborty et al. (2019) and Prompetchara et al. (2019) are a few great references in this regard. Even within time series techniques, autoregressive (AR) methods are found to be relevant and providing strong accuracy in multiple researches. More on this will follow in the next section.

Another common approach specifically aimed at modelling the peak counts or outbreaks, is known as epidemic modelling which generally adopts different variations of the susceptible–infected–recuperated (SIR) modelling approach. Most of the recent studies have explored and recommended the usage of fractional order differential equations for dengue modelling (Hamdan & Kilicman, 2019b).

Another elaborate fractional order modelling approach has been discussed in Al-Sulami et al. (2014) but was not tested on any dengue fever-related data. The same model was adopted and applied empirically on 2012 dengue epidemic data for Selangor, Malaysia by Hamdan and Kilicman (2019a). The results, however, were found to be overestimating the actual number of cases consistently, by both integer-order and fractional-order variants. Based on these published results, it can be argued that, while the approaches are theoretically sound, empirical results based on real data are at best volatile (Kilicman, 2018).

The third broad variant of dengue modelling is the regression-based approach. Halide and Ridd (2008) used stepwise regression to predict DHF cases for the city of Makassar and reported good performance in predicting moderately severe epidemic at lead times of up to 6 months. Current number of DHF cases and relative humidity were found to be the two most important predictors. It should, however, be noted that Makassar did not see any big outbreak during the study period, unlike San Juan or Iquitos. Bowman et al. (2016) found that recent probable cases, increase in mean temperature and mean age of hospitalized cases were adequately predictive of future counts at a lag of 1 to 12 weeks. Of these, the probable cases had the best performance, as captured using positive predictive values.

Apart from the above, machine learning approaches have been used in some related applications. The reader is referred to Guo et al. (2017) for a comparison of multiple advanced machine-learning algorithms such as support vector regression, gradient boosting regression trees etc. Another emerging technique in this context is to use ensemble modelling. Since different individual models work well under certain specific scenarios, an ensemble often provides more stable and superior results than any of its component models. One of the most popular ensemble techniques in the machine learning domain is the random forest (RF) model. It has been used for predicting disease outbreaks by many authors, and has been one of the more successful machine learning approaches in this regard. A few examples are Brasier et al. (2012), Eng et al. (2014) and Ong et al. (2018). On the other hand, ensemble method of analogous models, additive seasonal Holt-Winters and historical modelling, has shown strong results in forecasting peak height and total cases in a transmission season (Buczak et al., 2018). The same methodology was, however, not very strong in predicting the peak week precisely. Another interesting study was done by Guo et al. (2019), who used multiple regression models using different penalties, such as LASSO, Ridge, ElasticNet to form the ensemble. The resulting forecast provided superior results for 1–2 weeks lead time. Interestingly, this study used social media surveillance data to boost its prediction accuracy. A more detailed account of prior studies using ensemble approaches for disease outbreak forecasting can be found in Deb et al. (2017).

# 3 | METHODS

## 3.1 | Notations and definitions

Throughout this paper, wherever used, $(Y_t)$ denotes a univariate time series. $\mathcal{F}_t = \sigma(Y_i, i \leqslant t)$ denotes the sigma-field generated by the history of the series. Any boldfaced term, *e.g.* $\mathbf{w}$, is used to indicate a vector of appropriate order. For a set $S$, $|S|$ points to the cardinality of the set.

Akaike information criterion (AIC) is going to be used as a selection criteria in many cases. Recall that it is a combination of the likelihood and the number of parameters estimated in the model. A lower value of AIC is preferred.

For evaluating the predictions, we will use mean absolute error (MAE) throughout the study. Let $V$ be the prediction set, and for $t' \in V$, let $\widehat{Y}_{t'}$ be the prediction for true values $Y_{t'}$. MAE is then defined as

$$\text{MAE} = \frac{1}{|V|}\sum_{t' \in V}\left|Y_{t'} - \hat{Y}_{t'}\right|. \quad (1)$$

## 3.2 | Candidate models for ensembling

As mentioned in 1, our ensembling approach combines three different models, each coming with some advantages and some disadvantages. They are described below.

### 3.2.1 | Negative binomial regression

Using a negative binomial (NB) regression model is one of the most popular techniques for dealing with count data. This approach is considered to be efficient when the issue of overdispersion is observed in the count data, which is often the case for disease incidence data. Suppose, $\mathbf{x}_t$ is a column vector of covariates corresponding to $Y_t$, the number of dengue cases at the $t$th time point. Then, NB regression model assumes that $Y_t \sim \text{Poisson}(\mu_t)$ where the conditional mean parameter is defined as $\mu_t = \exp(\mathbf{x}_t^T\boldsymbol{\beta} + e_t)$, with $\boldsymbol{\beta}$ being the vector of unknown coefficients, and $e_t$ being a heterogeneity component unrelated to $\mathbf{x}_t$. Fairos et al. (2010) and Dhimal et al. (2015) are two of the many papers where NB regression has been used in the context of dengue forecasting. One of the best features of this method is its simplicity and ease of explainability. Having said that, performance of this model is generally poor when it encounters unusually high peak periods and is therefore not a great tool to predict epidemics. More on this will follow in the Results and the Discussion sections.

### 3.2.2 | Seasonal ARIMA

Autoregressive integrated moving average (ARIMA) methods are most popular for analysing continuous time series data. Box et al. (2015) is an excellent reading on this.

In the same spirit as above, consider a time series of dengue cases as $(Y_t)_{1\leqslant t\leqslant n}$ and assume it to be stationary. In case of nonstationarity, an initial differencing step is done. Then, an ARMA model of order $(p, q)$ is defined by

$$Y_t - \phi_1 Y_{t-1} - \ \dots\ - \phi_p Y_{t-p} = e_t + \theta_1 e_{t-1} + \ \dots\ + \theta_q e_{t-q}. \quad (2)$$

In the above, $e_i$'s are assumed to be independent and identically distributed Gaussian random variables with mean 0, the $\phi$ coefficients constitute the AR part and the $\theta$ coefficients form the moving average (MA) part. Furthermore, use $\phi(x) = 1 - \phi_1 x - \dots - \phi_p x^p$ to denote the AR characteristic polynomial, $\theta(x) = 1 + \theta_1 x + \dots + \theta_q x^q$ for the MA characteristic polynomial and $B$ to denote the backward shift operator. Then, the above model can be written in a simplified form as $\phi(B)Y_t = \theta(B)e_t$.

In practical applications, often the above model is extended to seasonal ARIMA (SARIMA) model to incorporate autocorrelation at seasonal lags. Suppose, $Y_t$ has a SARIMA structure with period $s$, AR order $p$, MA order $q$, seasonal AR order $p'$ and seasonal MA order $q'$. Let the seasonal AR coefficients be $\Phi_1, \dots, \Phi_{p'}$ and the seasonal MA coefficients be denoted by $\Theta_1, \dots, \Theta_{q'}$. Also, let $\phi(x)$ and $\theta(x)$ be defined as above, $\Phi(x) = 1 - \Phi_1 x^s - \dots - \Phi_{p'} x^{p's}$ and $\Theta(x) = 1 + \Theta_1 x^s + \dots + \Theta_{q'} x^{q's}$ be the characteristic polynomial corresponding to the seasonal components. Then, the following equation denotes the SARIMA model.

$$\Phi(B)\phi(B)Y_t = \Theta(B)\theta(B)e_t. \tag{3}$$

The above is often indicated as SARIMA$(p, d, q)(p', d', q')[s]$, where $d$ and $d'$ are the orders of differencing and seasonal differencing required to ensure stationarity for the data. For both ARIMA and SARIMA models, it is possible to include exogenous variables in the mean structure. It thus provides an attractive and effective framework to grasp the effects of various regressors and the linear trend while analysing a time series data.

SARIMA is the second model in our ensemble approach. Necessary differencing order to achieve stationarity is obtained from the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, cf. Kwiatkowski et al. (1992). Then, in order to choose the most appropriate orders, the AIC criterion is used. Maximum lags we consider for the AR and MA orders are 7.

On a related note, SARIMA models have been widely used for forecasting in the context of infectious diseases and other related areas. Some examples are tuberculosis (Rios et al., 2000), dengue fever (Luz et al., 2008), haemorrhagic fever with renal syndrome (Li et al., 2012), avian influenza H5N1 outbreaks (Kane et al., 2014) and influenza A (Petukhova et al., 2018). However, one major issue is that the discrete nature of the response variable may be a challenge while using it for count data.

### 3.2.3 | GLARMA

Generalized linear autoregressive moving average (GLARMA) method is an appropriate extension of the generalized linear model to deal with the serial dependence in the discrete time series data. They can be thought of as a class of observation-driven non-Gaussian nonlinear state-space models. Here, the state process is assumed to depend linearly on the exogenous variables whereas the dependence on past values of the process is assumed to be nonlinear. In this approach, the distribution of $Y_t$, which is a discrete variable, conditional on the history, is considered to be of the following exponential family form (Dunsmuir, 2015)

$$f(Y_t \mid \mathcal{F}_t) = \exp\left\{Y_t W_t - a_t b(W_t) + c_t\right\}. \tag{4}$$

Here, $W_t$ is a state variable defined as a function of the elements in $\mathcal{F}_t$ and $(a_t)$, $(c_t)$ denote sequences of constants. While there are a lot of choices for the specifications of $W_t$ (Benjamin et al., 2003), one of the most common form is the following (see Davis et al., 2003).

$$W_t = \mathbf{x}_t^T \boldsymbol{\beta} + Z_t,\ Z_t = \sum_{i=1}^{p} \phi_i (Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q} \theta_i e_{t-1}. \tag{5}$$

In the above equation, similar to before, $\mathbf{x}_t$ is the vector of covariates, and $\boldsymbol{\beta}$ is the coefficient vector. $Z_t$ is a noise process specified through an ARMA-type recursion. $(e_t)$ is the sequence of predictive residuals defined as $e_t = (Y_t - \mu_t)/\sigma_t$, where $\mu_t$ and $\sigma_t^2$ denote the conditional mean and variance of the time series. So far as the choice of the distribution goes, most common ones are Poisson, NB, and their generalized forms. Throughout this paper, we will work with Poisson GLARMA models, and for orders $p$,$q$ it will be denoted as GLARMA$(p, q)$. Detailed theoretical results in this aspect can be found in Davis et al. (1999), Benjamin et al. (2003), and Kedem and Fokianos (2005).

One of the main advantages of GLARMA models is that they deal with the discrete time series data in an appropriate way and can fit relatively easily to long time series. Because of that, GLARMA models have been applied in various disciplines. In one of the earliest applications, Rydberg and Shephard (2003) used it in the context of financial modelling. See Dunsmuir et al. (2008) and Etting and Isbell (2014) for a couple of other interesting applications. This framework has been explored in forecasting disease outbreak as well. Petukhova et al. (2018), for example, analysed the efficiency of GLARMA in predicting influenza A virus frequency.

## 3.3 | Ensemble method

In this work, a time-dependent weighting scheme has been used to generate ensemble forecasts. To illustrate the scheme, let $\widehat{Y}_t^{(1)}, \widehat{Y}_t^{(2)}, \widehat{Y}_t^{(3)}$ $(t \in U)$ indicate the fitted values by the three candidate models (refer to the previous section) for the train set $U$. Similarly, $\widehat{Y}_{t'}^{(1)}, \widehat{Y}_{t'}^{(2)}, \widehat{Y}_{t'}^{(3)}$ denote the predictions by the three models for $t' \in V$, where $V$ is the prediction set. Our aim is to predict $Y_{t'}, t \in V$, by a weighted combination of the form

$$\widehat{Y}_{t'} = \sum_{j=1}^{3} w_{t'}^{(j)} \widehat{Y}_{t'}^{(j)}. \tag{6}$$

We emphasize that the choice of the weights in the above equation is time dependent and that is where it becomes more attractive than common practices. In order to determine the appropriate weights, we look at the fitted data and compute the following, for $j = 1, 2, 3$ and for $t \in U$,

$$w_t^{(j)} = \exp\left\{ -\left| Y_t - \widehat{Y}_t^{(j)} \right| \right\} \Big/ \sum_{j=1}^{3} \exp\left\{ -\left| Y_t - \widehat{Y}_t^{(j)} \right| \right\}. \tag{7}$$

Thus, for every time point in the train set, the weights associated with an individual method are inversely proportional to the exponential function of the absolute values of the residuals. This is also known as the softmax function applied to the negative of the fitted residuals. Clearly, a method which does not fit the data well has less weight in the weighted average. Since this computation is done separately for every time point, the sequence $(w_t^{(j)})_{t \in U}$ captures the information of whether the $j$th model becomes more (or less) suitable for the data over time. Note the restriction $w_t^{(1)} + w_t^{(2)} + w_t^{(3)} = 1$. Furthermore, it is evident that the components of $(w_t^{(1)}, w_t^{(2)}, w_t^{(3)})$ are dependent among each other. They are autocorrelated as well. This motivates us to treat $(w_t^{(1)}, w_t^{(2)})_{t \in U \cup V}$ as a multivariate time series.

Defining $\text{logit}(x) = \log(x/(1-x))$, in order to obtain estimates of $w_{t'}^{(j)}$ for $t' \in V$, we fit a vector autoregressive (VAR) process for the bivariate time series $\mathbf{w}_t = (\text{logit}(w_t^{(1)}), \text{logit}(w_t^{(2)}))_{t \in U}$. Recall that a VAR model of lag $p$ is defined as

$$\mathbf{w}_t = \boldsymbol{\mu} + A_1 \mathbf{w}_{t-1} + A_2 \mathbf{w}_{t-2} + \ldots + A_p \mathbf{w}_{t-p} + \mathbf{e}_t. \tag{8}$$

Here, $\boldsymbol{\mu}$ is a bivariate constant vector, $A_i$'s are parameter matrices of appropriate order and $\mathbf{e}_t$ is a zero-mean error process with no temporal autocorrelation. The above model is used to forecast $\mathbf{w}_{t'}$ for $t' \in V$. Subsequently, Equation (6) provides the required forecast for the number of dengue cases.

## 3.4 | Forecast evaluation using model confidence set

Hansen et al. (2011) proposed the model confidence set (MCS) procedure which is used to test if one of the candidate models has superior predictive power over others. In order to describe the MCS procedure, let $\mathscr{M}$ be the set of models we consider and $V$ be the forecast period. We evaluate the forecast accuracy of the models under the MAE loss function. Let $L_{m,t}$ denote the loss associated with $m \in \mathscr{M}$, $t \in V$. The relative performance for $m_1, m_2 \in \mathscr{M}$ is then defined as $d_{m_1 m_2,t} = L_{m_1,t} - L_{m_2,t}$. Let $\mathbb{E}(d_{m_1 m_2,t})$ be the expectation of $d_{m_1 m_2,t}$ under the loss function. Then, the objective of the MCS procedure is to choose a model set $\mathscr{M}^*$ such that

$$\mathscr{M}^* = \left\{ m \in \mathscr{M} : \mathbb{E}(d_{mm',t}) \leqslant 0, \forall\ m' \in \mathscr{M} \right\}. \tag{9}$$

In order to do the above, the MCS procedure performs a sequence of significance tests with null hypothesis of the form

$$H_{0,\mathscr{M}'} : \mathbb{E}(d_{mm',t}) = 0 \ \forall\ m, m' \in \mathscr{M}' \subset \mathscr{M}. \tag{10}$$

In order to perform the above test, the authors suggested different choices. In this paper, we use the range statistic $\mathrm{T_R}$. If $\bar{d}_{mm'}$ and $v(d_{mm'})$ respectively denote the mean and variance of $d_{mm',t}$ for $t \in V$, then

$$\mathrm{T_R} = \max_{m,m' \in M'} \frac{\left|\bar{d}_{mm'}\right|}{\sqrt{v(d_{mm'})}}. \tag{11}$$

The asymptotic distribution of $\mathrm{T_R}$ is nonstandard, and it depends on some nuisance parameters. So, we use bootstrap methods to estimate the distribution and this approach implicitly solves the nuisance parameter problem, as pointed out by the authors in the original paper. Now, at every step of the hypothesis testing part, an equivalence test $\delta_{\mathscr{M}'}$ is conducted to test if any two models in $M'$ perform equally well under the loss function. If $\delta_{M'}$ is rejected, an elimination rule $e_{M'}$ eliminates the model with poor performance. These tests are repeated until the set of surviving models $\mathscr{M}^*$ is obtained. Throughout this paper, we use a significance level of 0.1 for the MCS procedure.

## 3.5 | Implementation

In the data we analyse, albeit no observation is missing for the response variable (the number of dengue cases), there are a few missing cases in the predictor variables. We choose to impute these missing values using the interpolation method, following Moritz and Bartz-Beielstein (2017). It is imperative to point out that the proportion of missing values is very less for most of the variables. Therefore, the imputation technique has little to no effect on the overall results. After the imputation, we focus on choosing the most appropriate set of terrain and climate related predictors and their lagged values. As Deb et al. (2017) have pointed out, the terrain and weather components from past weeks should be included in the model. In that light, throughout this study, we are going to use the covariate values from up to lag 4, that is previous 4 weeks. This is justified because the life cycle of *Aedes aegypti* can span from a few days to 3–4 weeks. However, note that this approach would effectively introduce a lot of new covariates, thereby increasing the chances of multicollinearity and overfitting. Thus, a variable selection step is warranted. That is achieved in two stages.

For the full set of predictors, the variance inflation factor criterion, with a cutoff of 10, is first used to remove the variables showing signs of multicollinearity. Next, the stepwise variable selection method is implemented on a usual logistic model with all predictors to choose the final set of climate- and terrain-related covariates. Additionally, to deal with seasonality, 11 monthly indicator variables (effect of December is taken as 0, for the sake of identifiability issues) are included as well. Note that for each of the candidate models (refer to Section 3.2), we use the same set of regressors.

AIC criterion is used to choose the most appropriate orders in SARIMA and GLARMA methods, and maximum lags of 7 are considered in this regard.

All computations are done in RStudio version 1.0.153, coupled with R version 3.5.2.

## 4 | RESULTS

As an application of the proposed method, we analyse the weekly number of dengue cases from two different regions. These data are publicly available in the GitHub repository DengAI (2017). The two regions are San Juan, Puerto Rico and Iquitos, Peru. For San Juan, the data spans from 30 April 1990 to 22 April 2008 whereas for the latter, the time-span is from 1 July 2000 to 25 June 2010.

In addition to the response variable which is the weekly number of dengue cases, we consider 13 main covariates in this study. They are NDVI or normalized difference vegetation index for four different directions; precipitation amount; reanalysis relative humidity and specific humidity; reanalysis air temperature, maximum air temperature and average air temperature; diurnal temperature range; average and maximum temperature of the stations. Summary statistics of all these variables are displayed in Table 1.

Note that the mean number of dengue cases for Iquitos and San Juan are 7.57 and 34.2 respectively. The corresponding ranges are [0, 116] and [0, 461]. These phenomena point to the fact that there are occasions of epidemics in both locations, as can be observed in Figure 1 as well.

The main objective of this work is to analyse and predict the number of dengue cases, especially the outbreaks, based on the aforementioned terrain and weather-related variables ahead of time. For that, we choose 6-week periods starting on different dates as different test sets. Then, for each of these test sets, the proposed model is trained on all data up to 8 weeks before the forecast period and we look at the predictive accuracy using MAE. So far as the method goes, missing data handling and variable selection are done as discussed in Section 3.5. The results are displayed in Figure 2.

The plot shows the average MAE for different 6-week periods in the forecast horizon starting on different dates. The results for San Juan are shown in the top panel and the same for Iquitos are presented in the bottom panel. Each 6-week period is categorized as high, medium or low, depending on the average number of observed cases. For San Juan, a high period is defined as a 6-week period with average number of cases greater than 40. This value is in the upper quartile, and signifies an epidemic period. A medium period is defined when the average number of cases is between 15 and 40 (which is around the median) and a low period is for the average cases being below 15 (in the lower quartile). In a similar fashion, noting that the quartiles of the total number of cases are much lower for Iquitos, high, medium and low period are defined for average number of cases being greater than 20, between 5 and 20 and less than 5 respectively. Then, for each of these cases, we take a look at the predictive accuracy if the prediction was made 1 week before and if it was made 8 weeks before. For example, the first point in the graph in the top panel depicts that the average MAE over a 6 week period starting on 26 October 2006 is 6.99 if the prediction was made on 19 October 2006. Similarly, if we consider the predictions made on 31 August 2006 (8 weeks before), the average MAE for the same period is 9.12.

**TABLE 1** Percentage of missing values, and the summary of the variables in the study. Data for San Juan spans from April 30 1990 to April 22 2008 whereas the timeline for the Iquitos data is from July 1 2000 to June 25 2010

| Summaries for San Juan | Missing values | Minimum | Maximum | Average | Std. dev. |
|---|---|---|---|---|---|
| Number of dengue cases | 0% | 0 | 461 | 34.2 | 51.38 |
| NDVI NE | 20.4% | −0.41 | 0.49 | 0.06 | 0.11 |
| NDVI NW | 5.2% | −0.46 | 0.44 | 0.07 | 0.09 |
| NDVI SE | 2.0% | −0.02 | 0.39 | 0.18 | 0.06 |
| NDVI SW | 2.0% | −0.06 | 0.38 | 0.17 | 0.06 |
| Precipitation amount (mm) | 1.0% | 0.00 | 390.60 | 35.47 | 44.61 |
| Reanalysis air temperature (K) | 0.6% | 295.94 | 302.20 | 299.16 | 1.24 |
| Reanalysis avg air temperature (K) | 0.6% | 296.11 | 302.16 | 299.28 | 1.22 |
| Reanalysis max air temperature (K) | 0.6% | 297.80 | 304.30 | 301.40 | 1.26 |
| Reanalysis relative humidity (%) | 0.6% | 66.74 | 87.58 | 78.57 | 3.39 |
| Reanalysis specific humidity (g/kg) | 0.6% | 11.72 | 19.44 | 16.55 | 1.56 |
| Reanalysis diurnal temp range (K) | 0.6% | 1.36 | 4.43 | 2.52 | 0.50 |
| Station avg temperature (C) | 0.6% | 22.84 | 30.07 | 27.01 | 1.42 |
| Station max temperature (C) | 0.6% | 26.70 | 35.60 | 31.61 | 1.72 |
| **Summaries for Iquitos** | **Missing values** | **Minimum** | **Maximum** | **Average** | **Std. dev.** |
| Number of dengue cases | 0% | 0 | 116 | 7.57 | 10.77 |
| NDVI NE | 0.6% | 0.06 | 0.51 | 0.26 | 0.08 |
| NDVI NW | 0.6% | 0.04 | 0.45 | 0.24 | 0.08 |
| NDVI SE | 0.6% | 0.03 | 0.54 | 0.25 | 0.08 |
| NDVI SW | 0.6% | 0.06 | 0.55 | 0.27 | 0.09 |
| Precipitation amount (mm) | 0.8% | 0.00 | 210.83 | 64.25 | 35.22 |
| Reanalysis relative humidity (%) | 0.8% | 57.79 | 98.61 | 88.64 | 7.58 |
| Reanalysis air temperature (K) | 0.8% | 294.64 | 301.64 | 297.87 | 1.17 |
| Reanalysis avg air temperature (K) | 0.8% | 294.89 | 302.93 | 299.13 | 1.33 |
| Reanalysis max air temperature (K) | 0.8% | 300.00 | 314.00 | 307.08 | 2.38 |
| Reanalysis specific humidity (g/kg) | 0.8% | 12.11 | 20.46 | 17.10 | 1.45 |
| Reanalysis diurnal temp range (K) | 0.8% | 3.71 | 16.03 | 9.21 | 2.45 |
| Station avg temperature (C) | 7.1% | 21.40 | 30.80 | 27.53 | 0.92 |
| Station max temperature (C) | 2.7% | 30.10 | 42.20 | 34.00 | 1.33 |

From the top panel of the graph (results for San Juan), we can see that the average MAE goes above 10 for the ensemble method only in two cases, for the 6-week period starting on 8 November 2007 and on 20 December 2007. First one is a medium period (average number of cases is 21.67) while the latter one is a low period (average number of cases is 13.33). Other than these two cases, the proposed method has been very accurate. Even when the predictions are made 8 weeks in advance, the MAE does not go above 10.

Results for Iquitos are displayed in the bottom panel of Figure 2. Once again, only for the first case (6-week period starting from 30 September 2008) which is in fact a high period, the MAE is above 20 when we look at the predictions made 8 weeks before. The MAE for the same period according to predictions made 1 week before falls below 20. Other than that, the MAE remains below 10 throughout.

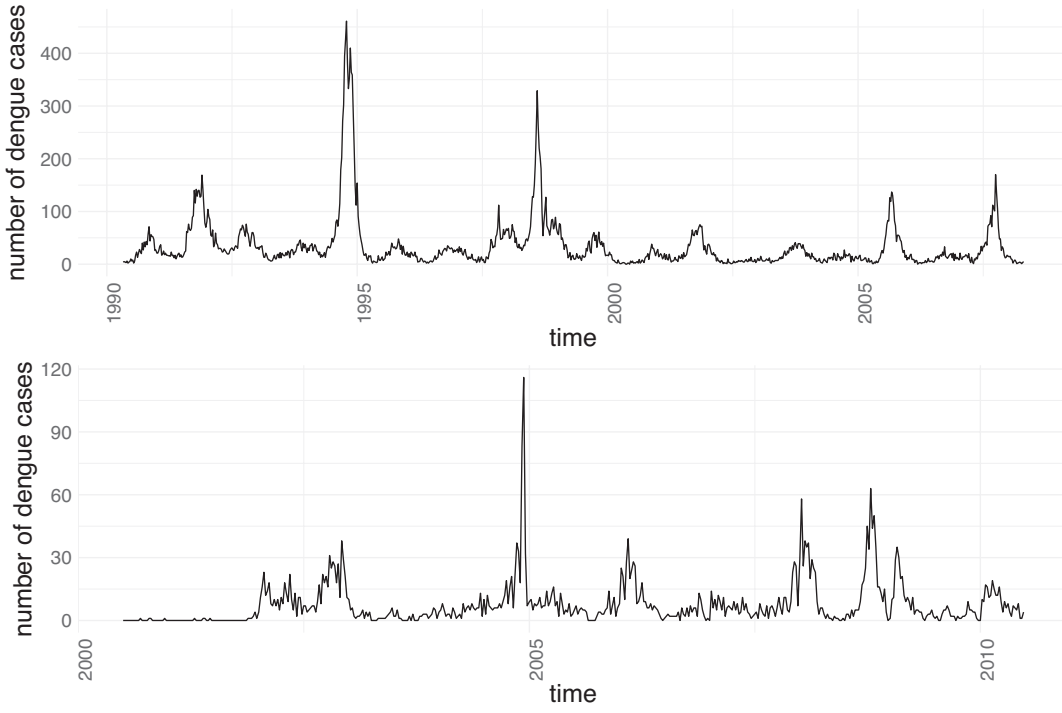


**FIGURE 1** Time series of the weekly number of dengue cases for San Juan (in top panel) and Iquitos (in bottom panel)

We see that the predictions from 8 weeks before ensure good accuracy and it firmly establishes that using appropriate lagged values of climate and terrain-related covariates, early prediction of dengue outbreak is possible.

Next, we turn our attention to the early prediction part and for the same periods above, compute the predictive accuracy of the main model and the three individual models if the forecast was done 8 weeks before. In this comparative study, we also include the results from another ensemble model, namely the RF. Recall that the RF is arguably the most popular ensemble method in related applications and can be taken as a benchmark. In order to keep uniformity in the analysis, the RF model is implemented with the same set of regressors and the hyperparameters are tuned appropriately. Now, MCS procedure as described in Section 3.4 is performed on the eight-weeks-ago forecasts from the five models (RF, NB, ARIMA, GLARMA and ensemble). The $p$-values from the MCS procedure are displayed in Figure 3. Note that a value greater than 0.1 (dotted line in the graph) indicates an acceptable model whereas a $p$-value of 1 corresponds to the best model in terms of prediction.

From the graphs, it is evident that the ensemble method outperforms the other candidate models for all of the high periods (three for San Juan and one for Iquitos). Among the medium periods, there is one in San Juan (starting from 26 October 2006) and two in Iquitos (starting from 11 November 2008 and 16 February 2010) where the ensemble method is bettered by the RF model. Otherwise, the proposed method is the best. However, for the low periods, there is no consistent pattern. Overall, we can say that all the models are equally accurate to predict the low numbers of dengue cases whereas for early prediction of dengue outbreak, our proposed method usually outperforms other candidate models, including the RF which in itself is another type of ensemble method.

Furthermore, we take a more detailed look at three specific examples of the above 6-week periods—one high, one medium and one low—for both locations. For San Juan, we choose the 6-week windows

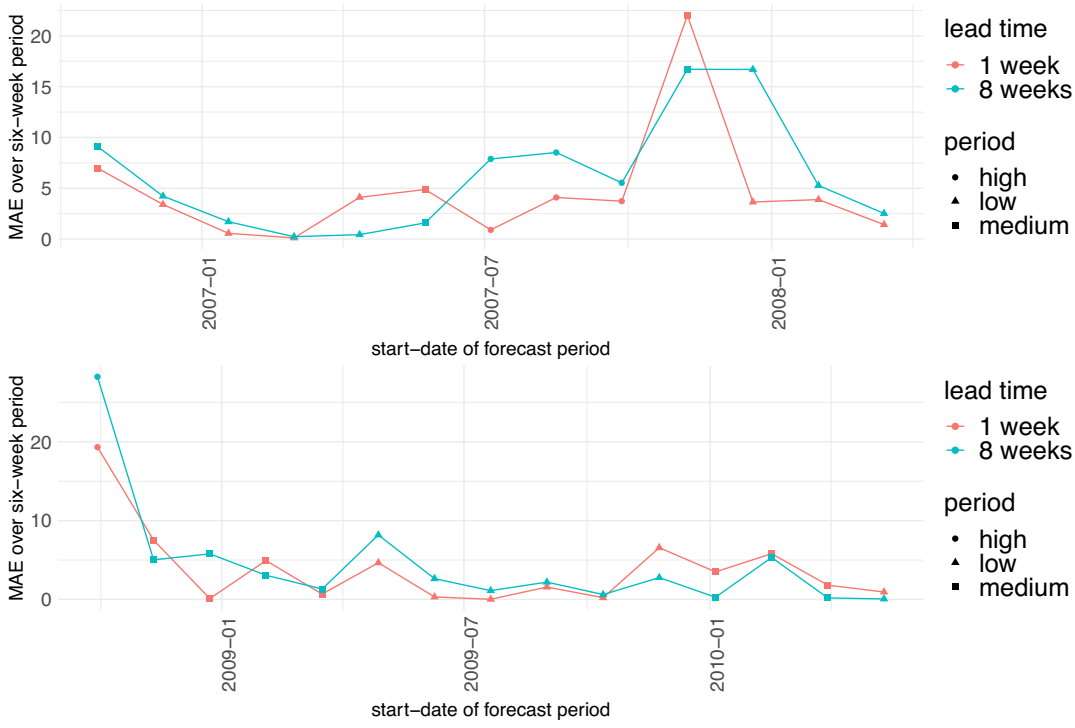


**FIGURE 2** Average 6-week MAE of predictions made 1 week and 8 weeks ahead of time for forecast periods starting on different dates. High, medium and low periods are shown using different symbols. Results for San Juan are shown in the top panel and the same for Iquitos are presented in the bottom panel [Colour figure can be viewed at wileyonlinelibrary.com]

starting from 20 August 2007 (high period), 18 June 2007 (medium period) and 11 March 2008 (low period). For Iquitos, the same dates are chosen as 1 October 2008, 12 February 2010 and 21 May 2010 respectively. Here we consider different scenarios where all data up to $d$ weeks (for $d = 1, 2, 3, 4, 6, 8$) before the forecast period are used to train the model and then we get predictions for the forecast period. This is done to understand the robustness of forecast accuracy corresponding to different lead times. The mean absolute error for the predictions in the high, medium and low periods for San Juan and Iquitos is displayed in Tables 2 and 3.

Interestingly, for San Juan, we can see that the ensemble method performs way better than the other models. For the high period, the average number of cases is 93, and the MAE for all lead times is below 10 for the proposed method. In comparison, the other models observe errors at least four times in magnitude. For the medium period (average cases 34.67), once again, the ensemble method beats the other models whereas the RF is close second for most of the lead times. On the other hand, all models perform at par with each other for the low period (average cases 2.67).

Superior results for the ensemble method can be noted for the high period in Iquitos as well, except when the predictions are made 1 week or 2 weeks before. In that case, RF beats the proposed method. However, the magnitude of difference in the MAE is much smaller here. Meanwhile, the NB and the RF models are the best models for the medium period while the ensemble and ARIMA are better than others for the low period. Further note that the best MAE values are very small in all these cases.

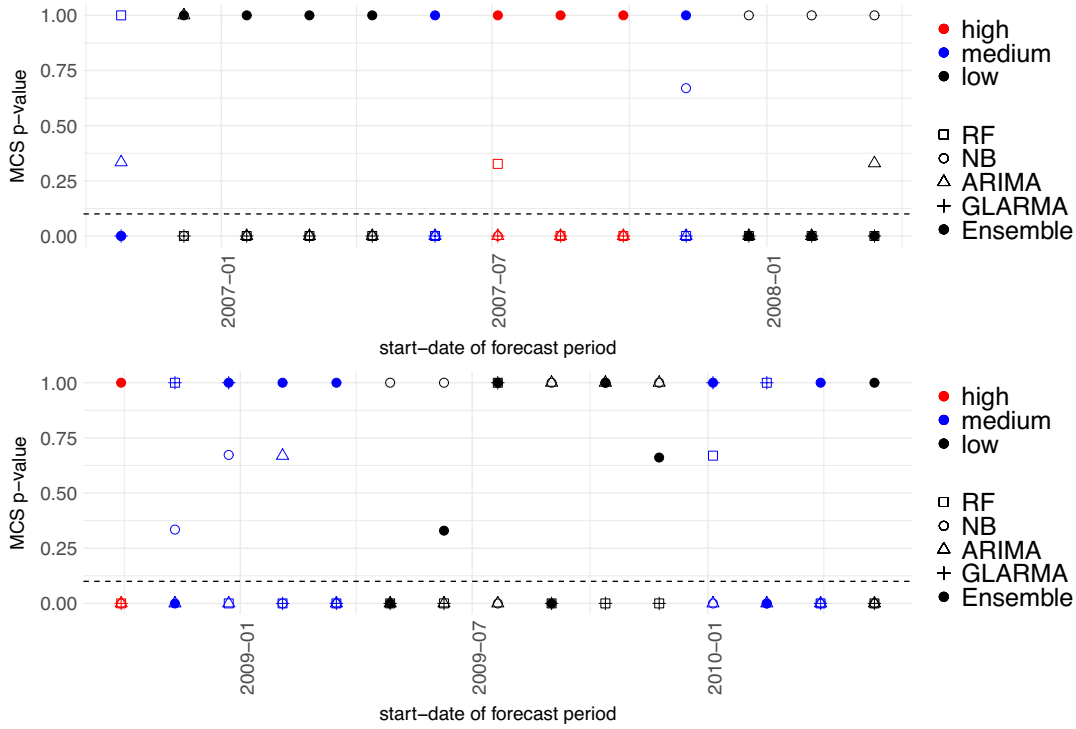


**FIGURE 3** *p*-values from the MCS procedure for forecasts made 8 weeks in advance. Results for San Juan are shown in the top panel and the same for Iquitos are presented in the bottom panel [Colour figure can be viewed at wileyonlinelibrary.com]

Next, we look at the precision and recall for the 8 weeks ago predictions of the proposed method for all of the high periods within the last 5 years of the corresponding dataset. These results are presented in Table 4. Keeping the same notion as with earlier analysis, a high or epidemic period is defined as a 6-week period where the average number of cases is greater than 40 for San Juan and above 20 for Iquitos.

There are 35 high periods for San Juan, and the ensemble method records great precision and recall. For Iquitos though, the values are a little bit smaller, perhaps owing to the fact that the size of the training data is less in this case.

As a final piece of this study, we look at the robustness of the proposed ensemble method in terms of the length and resolution of the training data. For brevity, all of the remaining results presented in this section are based on 8 weeks ago predictions, for that is the more practical aspect of the problem. We, however, emphasize that changing the lead time does not alter the results significantly.

First, in an attempt to understand the effect of the size of the training set, we focus on all of the forecast periods discussed in Figure 2, but use only the last $k$ years of data (for $k = 2, 3, 5, 8$) to make predictions. Then, the average mean absolute errors are calculated for each case and are presented in Figure 4. With 2 years of data, the performance is somewhat abrupt, potentially due to the method's inability to extract sufficient information about the seasonality in the time series. Results for San Juan (top panel of the figure) indicate that the errors are less than 20 in all but a couple of forecast periods. Those two periods (around August–September of 2007) are epidemic periods and saw more than 90 cases on an average. The ensemble method, when only up to 5 years of data are used to train

**TABLE 2** MAE for predictions in three different periods, corresponding to different lead times by candidate models for San Juan data. High period starts on 20 August 2007 and the average number of dengue cases is 93. Medium period, with average number of dengue cases as 34.67, starts on 18 June 2007. The low period starts on 11 March 2008 and the average number of cases is 2.67. Best models are marked in bold

| Period | Lead-time | RF | NB | ARIMA | GLARMA | Ensemble |
|---|---|---|---|---|---|---|
| High | 1 week | 16.83 | 89.31 | 26.33 | 42.50 | **4.49** |
| | 2 weeks | 29.33 | 89.21 | 32.88 | 38.65 | **6.71** |
| | 3 weeks | 45.50 | 89.07 | 35.41 | 47.34 | **8.67** |
| | 4 weeks | 49.33 | 89.27 | 32.14 | 51.82 | **6.62** |
| | 6 weeks | 41.83 | 89.71 | 32.42 | 61.15 | **9.69** |
| | 8 weeks | 46.33 | 89.08 | 31.87 | 45.68 | **8.51** |
| Medium | 1 week | 6.33 | 32.07 | 9.52 | 15.48 | **4.24** |
| | 2 weeks | 7.50 | 31.54 | 10.78 | 15.13 | **5.10** |
| | 3 weeks | 8.00 | 32.03 | 17.32 | 18.91 | **6.33** |
| | 4 weeks | 10.50 | 31.62 | 19.83 | 16.43 | **5.70** |
| | 6 weeks | 13.50 | 31.91 | 25.17 | 16.65 | **5.04** |
| | 8 weeks | 7.33 | 31.76 | 27.40 | 18.43 | **4.93** |
| Low | 1 week | 3.50 | **1.18** | 1.76 | 4.13 | 1.47 |
| | 2 weeks | 3.00 | **0.97** | 1.99 | 6.67 | 2.09 |
| | 3 weeks | 3.17 | **0.94** | 1.36 | 6.77 | 2.17 |
| | 4 weeks | 3.67 | **1.00** | 2.39 | 6.89 | 1.94 |
| | 6 weeks | 4.83 | **0.81** | 2.06 | 8.67 | 3.14 |
| | 8 weeks | 4.50 | **0.98** | 4.21 | 7.36 | 3.12 |

the model, records an MAE of around 40. We also notice that the error decreases with bigger training sets and earlier we saw that the error is less than 10 when all data are used in the model. For all other periods in San Juan, even with only 3 years of data, the method achieves great accuracy. A similar phenomenon is observed in all periods of Iquitos as well. Comparing the bottom panels of Figures 2 and 4, it is easy to conclude that the size of the training set has very little effect on the prediction accuracy of the ensemble method in the case of Iquitos. Such performance stability can be attributed to the presence of high and medium periods within the extracted training sets. As long as the training set covers a number of such periods, the ensemble model is able to learn and provide accurate estimates even with small training sets. In summary, the method appears to be performing well even when the time series is shorter.

Next, we check the performance of the model if the data are aggregated to a less granular level. Here, we transform the weekly number of cases to the monthly number of cases, by taking the average of all the values within a particular month. This is done to keep in line with the earlier analysis. It results in a time series of length 217 for San Juan and a time series of length 120 for Iquitos. In both cases, we focus on the last 4 months of data, and find out the accuracy of the 1 month ago or the 2 months ago predictions (equivalent to 8 weeks ago predictions). These results are provided in Table 5.

We can see that the errors in predicting the average number of cases are still less than 10 barring a couple of forecast periods in San Juan. In those cases, the proposed method records considerably higher errors. We hypothesize that the monthly or more coarse aggregations average the fluctuations in both the response and the predictor variables and as a result, the information content and

**TABLE 3** MAE for predictions in three different periods, corresponding to different lead times by candidate models for Iquitos data. High period starts on 30 September 2008 and the average number of dengue cases is 40.33. Medium period, with average number of dengue cases as 14.33, starts on 11 February 2010. The low period starts on 20 May 2010 and the average number of cases is 4.17. Best models are marked in bold

| Period | Lead time | RF | NB | ARIMA | GLARMA | Ensemble |
|---|---|---|---|---|---|---|
| High | 1 week | **17.67** | 28.36 | 18.04 | 37.89 | 19.33 |
| | 2 weeks | **21.83** | 29.91 | 24.91 | 38.01 | 22.27 |
| | 3 weeks | 25.33 | 30.47 | 23.95 | 38.40 | **22.53** |
| | 4 weeks | 26.50 | 31.36 | 31.09 | 38.38 | **25.13** |
| | 6 weeks | 29.83 | 31.52 | 36.21 | 38.85 | **27.34** |
| | 8 weeks | 33.50 | 32.32 | 37.74 | 38.10 | **28.23** |
| Medium | 1 week | 2.33 | **1.89** | 9.30 | 12.07 | 4.72 |
| | 2 weeks | **2.17** | 4.92 | 12.60 | 12.29 | 7.49 |
| | 3 weeks | **3.00** | 5.19 | 13.88 | 12.46 | 8.05 |
| | 4 weeks | 5.17 | **4.16** | 14.18 | 12.52 | 7.38 |
| | 6 weeks | 6.00 | **3.17** | 13.94 | 11.81 | 6.22 |
| | 8 weeks | **2.17** | 3.53 | 11.02 | 11.92 | 5.63 |
| Low | 1 week | 5.33 | 1.76 | 0.89 | 2.73 | **0.15** |
| | 2 weeks | 7.17 | 2.30 | 2.46 | 3.20 | **0.82** |
| | 3 weeks | 8.00 | 4.67 | **0.56** | 2.82 | 0.89 |
| | 4 weeks | 9.67 | 2.07 | 3.22 | 3.27 | **1.10** |
| | 6 weeks | 11.00 | 3.15 | 3.71 | 3.33 | **1.03** |
| | 8 weeks | 11.33 | 5.30 | 1.89 | 2.33 | **1.87** |

**TABLE 4** Precision and Recall for 8 weeks ago predictions of the ensemble method for high/epidemic periods

| Location | San Juan | Iquitos |
|---|---|---|
| Test period | 1 January 2004 to 22 April 2008 | 1 January 2006 to 14 May 2010 |
| No of high periods | 35 | 29 |
| Precision | 84.62% | 42.11% |
| Recall | 94.29% | 27.59% |

the predictive power of the independent variables go down. This impacts the prediction accuracy adversely, especially in case of high or medium periods. However, if the monthly data are available for a longer time period and if the data include sufficient number of high and medium periods, the ensemble approach should be able to achieve greater accuracy.

## 5 | DISCUSSION

In this study, we have proposed a new method to analyse and forecast time series count data in the context of dengue outbreaks. The ensemble approach is precise and easy to interpret. The VAR method of estimating the weights enables us to leverage the efficacy of every method in the system.

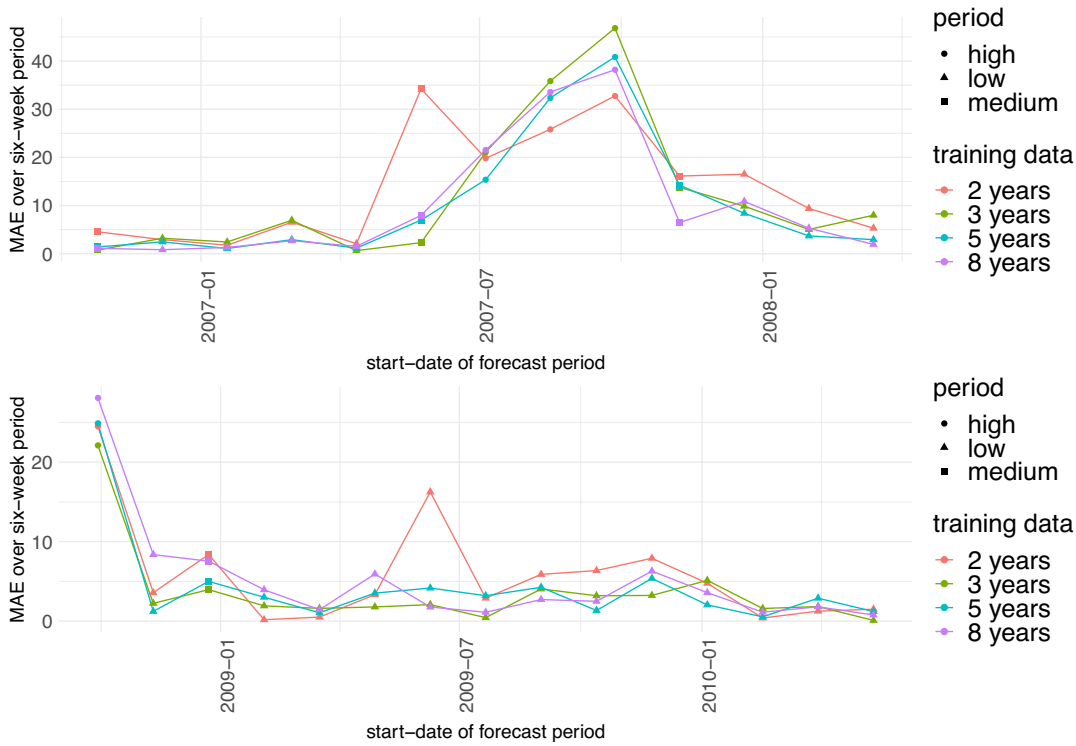


**FIGURE 4** Average 6-week MAE of predictions made 8 weeks ahead of time for forecast periods starting on different dates, corresponding to training data of different sizes. High, medium and low periods are shown using different symbols. Results for San Juan are shown in the top panel and the same for Iquitos are presented in the bottom panel [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 5** MAE of the 1 month ago and the two months ago predictions of the ensemble method when monthly data are considered

| Location | Month | Period | MAE (one month ago) | MAE (two months ago) |
|---|---|---|---|---|
| San Juan | November, 2007 | Medium | 42.0 | 31.7 |
| | December, 2007 | Low | 10.0 | 29.7 |
| | January, 2008 | Low | 0.1 | 5.2 |
| | February, 2008 | Low | 3.1 | 2.2 |
| | March, 2008 | Low | 1.4 | 2.5 |
| | April, 2008 | Low | 0.3 | 0.2 |
| Iquitos | January, 2010 | Medium | 4.3 | 0.2 |
| | February, 2010 | Medium | 5.8 | 7.0 |
| | March, 2010 | Medium | 7.3 | 7.0 |
| | April, 2010 | Medium | 1.7 | 4.1 |
| | May, 2010 | Low | 6.4 | 8.8 |
| | June, 2010 | Low | 1.3 | 3.8 |

It is flexible in nature. So, although we include three different models in the ensemble, one can easily include more types of models that may deem suitable for the problem. More regressors and their lagged values can also be included in the mean structure as necessary.

The real-life examples from San Juan and Iquitos show that the method works well to predict dengue epidemics at least 8 weeks in advance. It is considerably better than other competing models in that aspect. In comparison, the forecast accuracy is similar when the target period has a lesser number of cases. A couple of interesting phenomena are worth mention at this point. Akin to the existing literature which establish that the NB regression model suffers from the presence of outliers, it is observed that the NB model fails miserably in the case of high periods in our data too. It, however, performs very well for the low periods (e.g. it is the single best model for all lead times in case of the discussed low period in San Juan), thereby motivating its inclusion in the ensemble. Furthermore, the structure of the NB model does not capture the time-dependent nature of the data and that is also responsible behind its poor performance in case of the high periods. This is where the other two models perform better. If there is a series of weeks with an increasing number of cases, ARIMA or GLARMA can leverage that to provide considerably better forecasts. This in turn helps the ensemble to achieve great accuracy as well. Another critical observation is that the RF model has a tendency to over-forecast the low periods and under-forecast the high periods. It is actually a common problem with this approach. Sometimes it cannot analyse the relationships between the response and the predictors efficiently, and as a consequence, the predictions tend to move towards the average.

The results also establish that climate and terrain covariates from past weeks are instrumental in predicting the outbreaks. On a related note, the results from San Juan are much better than that from Iquitos. This can be attributed to the fact that San Juan has recorded more dengue epidemics over the years and the pattern can be successfully captured by our proposed method.

It is clear that it is of significant importance for healthcare as well as public health governing bodies to detect ensuing dengue epidemics. The method described in this paper can do that with strong confidence and with a reasonable lead time, thereby helping both health systems and disease control programs to respond and plan both preventive and curative measures more efficiently.

## DECLARATION OF INTEREST

The authors declare no conflict of interest.


## ACKNOWLEDGEMENTS

The authors thank two anonymous referees and the editor for their constructive comments which have improved the paper significantly.



## ORCID

*Soudeep Deb* https://orcid.org/0000-0003-0567-7339


## REFERENCES


Ahmed, N., Rafiq, M., Baleanu, D., Alshomrani, A.S. & Rehman, M.A.U. (2020) Positive explicit and implicit computational techniques for reaction–diffusion epidemic model of dengue disease dynamics. *Advances in Difference Equations*, 2020, 1–22. https://doi.org/10.1186/s13662-020-02622-z

Al-Sulami, H., El-Shahed, M., Nieto, J.J. & Shammakh, W. (2014) On fractional order Dengue epidemic model. *Mathematical Problems in Engineering* 2014. https://doi.org/10.1155/2014/456537

Benjamin, M.A., Rigby, R.A. & Stasinopoulos, D.M. (2003) Generalized autoregressive moving average models. *Journal of the American Statistical association*, 98, 214–223. https://doi.org/10.1198/016214503388619238

Bowman, L.R., Tejeda, G.S., Coelho, G.E., Sulaiman, L.H., Gill, B.S., McCall, P.J. et al. (2016) Alarm variables for Dengue outbreaks: A multi-centre study in Asia and Latin America. *PLoS One*, 11, e0157971. https://doi.org/10.1371/journal.pone.0157971

Box, G.E., Jenkins, G.M., Reinsel, G.C. & Ljung, G.M. (2015) *Time series analysis: Forecasting and control*. Hoboken: John Wiley & Sons.

Brasier, A.R., Ju, H., Garcia, J., Spratt, H.M., Victor, S.S., Forshey, B.M. et al. (2012) A three-component biomarker panel for prediction of Dengue Hemorrhagic Fever. *The American Journal of Tropical Medicine and Hygiene*, 86, 341–348. https://doi.org/10.4269/ajtmh.2012.11-0469

Buczak, A.L., Baugher, B., Moniz, L.J., Bagley, T., Babin, S.M. & Guven, E. (2018) Ensemble method for dengue prediction. *PLoS One*, 13 https://doi.org/10.1371/journal.pone.0189988

Chakraborty, T., Chattopadhyay, S. & Ghosh, I. (2019) Forecasting Dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications*, 527, 121266. https://doi.org/10.1016/j.physa.2019.121266

Davis, R.A., Dunsmuir, W.T. & Wang, Y. (1999) Modeling time series of count data. *Statistics Textbooks and Monographs*, 158, 63–114.

Davis, R.A., Dunsmuir, W.T. & Streett, S.B. (2003) Observation-driven models for Poisson counts. *Biometrika*, 90, 777–790. https://doi.org/10.1093/biomet/90.4.777

Deb, S., Acebedo, C.M.L., Dhanapal, G. & Heng, C.M.C. (2017) An ensemble prediction approach to weekly Dengue cases forecasting based on climatic and terrain conditions. *Journal of Health and Social Sciences* 2, 257–272. https://doi.org/10.19204/2017/nnsm3

DengAI, D. (2017) DengAI: predicting disease spread—a competition hosted by drivendata. https://github.com/ngbolin/DengAI

Dhimal, M., Gautam, I., Joshi, H.D., O'Hara, R.B., Ahrens, B., Kuch, U. (2015) Risk factors for the presence of chikungunya and dengue vectors (Aedes aegypti and Aedes albopictus), their altitudinal distribution and climatic determinants of their abundance in central nepal. *PLoS neglected tropical diseases*, 9, e0003545. https://doi.org/10.1371/journal.pntd.0003545

Dom, N.C., Hassan, A.A., AbdLatif, Z., Ismail, R., (2013) Generating temporal model using climate variables for the prediction of Dengue cases in Subang Jaya, Malaysia. *Asian Pacific journal of tropical disease*, 3, 352–361. https://doi.org/10.1016/S2222-1808(13)60084-5

Dunsmuir, W.T. (2015) Generalized linear autoregressive moving average models. Handbook of Discrete-Valued Time Series. CRC Monographs.

Dunsmuir, W., Tran, C.D., Weatherburn, D. & Wales, N. (2008) *Assessing the impact of mandatory DNA testing of prison inmates in NSW on clearance, charge and conviction rates for selected crime categories.* Australia NSW Bureau of Crime Statistics and Research Sydney.

Ebi, K.L. & Nealon, J. (2016) Dengue in a changing climate. *Environmental research*, 151, 115–123. https://doi.org/10.1016/j.envres.2016.07.026

Eng, C.L., Tong, J.C. & Tan, T.W. (2014) Predicting host tropism of influenza A virus proteins using random forest. *BMC medical genomics*, 7, S1. https://doi.org/10.1186/1755-8794-7-S3-S1

Estallo, E.L., Luduena-Almeida, F.F., Visintin, A.M., Scavuzzo, C.M., Lamfri, M.A., Introini, M.V. et al. (2012) Effectiveness of normalized difference water index in modelling Aedes ae-gypti house index. *International journal of remote sensing*, 33, 4254–4265. https://doi.org/10.1080/01431161.2011.640962

Etting, S.F. & Isbell, L.A. (2014) Rhesus macaques (Macaca mulatta) use posture to assess level of threat from snakes. *Ethology*, 120, 1177–1184. https://doi.org/10.1111/eth.12293

Fairos, W.W., Azaki, W.W., Alias, L.M. & Wah, Y.B. (2010) Modelling dengue fever (DF) and dengue haemorrhagic fever (DHF) outbreak using Poisson and Negative Binomial model. *International Journal Mathematics Computer Science Engineering*, 4, 809–814. https://publications.waset.org/vol/38

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q. et al. (2017) Developing a dengue forecast model using machine learning: A case study in China. *PLoS Neglected Tropical Diseases*, 11, e0005973. https://doi.org/10.1371/journal.pntd.0005973

Guo, P., Zhang, Q., Chen, Y., Xiao, J., He, J., Zhang, Y. et al. (2019) An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. *Science of The Total Environment*, 647, 752–762. https://doi.org/10.1016/j.scitotenv.2018.08.044

Halide, H. & Ridd, P. (2008) A predictive model for Dengue Hemorrhagic Fever epidemics. *International Journal of Environmental Health Research*, 18, 253–265. https://doi.org/10.1080/09603120801966043

Hamdan, N., & Kilicman, A. (2019a) Analysis of the fractional order dengue transmission model: a case study in Malaysia. *Advances in Difference Equations*, 2019, 31. https://doi.org/10.1186/s13662-019-1981-z

Hamdan, N. & Kilicman, A. (2019b) Basic epidemic model of Dengue transmission using the fractional order differential equations. *MJS*, 38, 1–18. https://doi.org/10.22452/mjs.sp2019no1.1

Hansen, P.R., Lunde, A. & Nason, J.M. (2011) The model confidence set. *Econometrica*, 79, 453–497. https://doi.org/10.3982/ECTA5771

Kane, M.J., Price, N., Scotch, M. & Rabinowitz, P. (2014) Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, 15, 276. https://doi.org/10.1186/1471-2105-15-276

Kearney, M., Porter, W.P., Williams, C., Ritchie, S. & Hoffmann, A.A. (2009) Integrating biophysical models and evolutionary theory to predict climatic impacts on species' ranges: the dengue mosquito Aedes aegypti in Australia. *Functional Ecology*, 23, 528–538. https://doi.org/10.1111/j.1365-2435.2008.01538.x

Kedem, B. & Fokianos, K. (2005) *Regression models for time series analysis*, Volume 488. Hoboken: John Wiley & Sons.

Kilicman, A. (2018) A fractional order SIR epidemic model for Dengue transmission. *Chaos, Solitons & Fractals*, 114, 55–62. https://doi.org/10.1016/j.chaos.2018.06.031

Kwiatkowski, D., Phillips, P.C., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54, 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Li, Q., Guo, N.N., Han, Z.Y., Zhang, Y.B., Qi, S.X., Xu, Y.G. et al. 2012. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. *The American journal of tropical medicine and hygiene*, 87, 364–370. https://doi.org/10.4269/ajtmh.2012.11-0472

Luz, P.M., Mendes, B.V., Codeço, C.T., Struchiner, C.J. & Galvani, A.P. (2008) Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *The American Journal of Tropical Medicine and Hygiene*, 79, 933–939. https://doi.org/10.4269/ajtmh.2008.79.933

Moritz, S. & Bartz-Beielstein, T. (2017) imputeTS: Time series missing value imputation in R. *R Journal*, 9, 207.

Ong, J., Liu, X., Rajarethinam, J., Kok, S.Y., Liang, S., Tang, C.S. et al. (2018) Mapping dengue risk in Singapore using random forest. *PLoS neglected tropical diseases*, 12, e0006587. https://doi.org/10.1371/journal.pntd.0006587

Petukhova, T., Ojkic, D., McEwen, B., Deardon, R. & Poljak, Z. (2018) Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. *PLoS One* 13, e0198313. https://doi.org/10.1371/journal.pone.0198313

Prompetchara, E., Ketloy, C., Thomas, S.J. & Ruxrungtham, K. (2019) Dengue vaccine: global development update. *Asian Pacific Journal of Allergy and Immunology*, https://doi.org/10. https://doi.org/10.12932/AP-100518-0309

Rios, M., Garcia, J., Sanchez, J. & Perez, D. (2000) A statistical analysis of the seasonality in pulmonary tuberculosis. *European Journal of Epidemiology*, 16, 483–488. https://doi.org/10.1023/a:1007653329972

Rydberg, T.H. & Shephard, N. (2003) Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics*, 1, 2–25. https://doi.org/10.1093/jjfinec/nbg002

Wu, P.C., Guo, H.R., Lung, S.C., Lin, C.Y. & Su, H.J. (2007) Weather as an effective predictor for occurrence of dengue fever in Taiwan. *Acta Tropica*, https://doi.org/103, 50–57.

Yang, H., Macoris, M.D.L.D.G., Galvani, K., Andrighetti, M. & Wanderley, D. (2009) Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. *Epidemiology & Infection*, 137, 1188–1202. https://doi.org/10.1017/S0950268809002040

**How to cite this article:** Deb S, Deb S. An ensemble method for early prediction of dengue outbreak. *J R Stat Soc Series A*. 2022;185:84–101. https://doi.org/10.1111/rssa.12714