Regular Article

When climate variables improve the dengue forecasting: a machine learning approach

Sidney T. da Silva[1], Enrique C. Gabrick[2,3,4,a], Paulo R. Protachevicz[2,5], Kelly C. Iarosz[4,6], Iberê L. Caldas[5], Antonio M. Batista[4,7], and Jürgen Kurths[2,3]

[1] Federal University of Paraná, Curitiba, PR 81531-980, Brazil
[2] Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany
[3] Department of Physics, Humboldt University Berlin, Newtonstraße 15, 12489 Berlin, Germany
[4] Graduate Program in Science, State University of Ponta Grossa, Ponta Grossa, PR 84030-900, Brazil
[5] Institute of Physics, University of São Paulo, São Paulo, SP 05508-090, Brazil
[6] University Center UNIFATEB, Telêmaco Borba, PR 84266-010, Brazil
[7] Department of Mathematics and Statistics, State University of Ponta Grossa, Ponta Grossa, PR 84030-900, Brazil

Abstract Dengue is a viral vector-borne infectious disease that affects many countries worldwide, infecting around 390 million people per year. The main outbreaks occur in subtropical and tropical countries. We, therefore, study here the influence of climate on dengue. In particular, we consider dengue and meteorological data from Natal (2016–2019), Brazil, Iquitos (2001–2012), Peru, and Barranquilla (2011–2016), Colombia. For the analysis and simulations, we apply machine learning (ML) techniques, especially the random forest (RF) algorithm. We utilize dengue disease cases and climate data delayed by up to one week to forecast the cases of dengue. In addition, regarding as feature in the ML technique, we analyze three possibilities: only dengue cases (D); climate and dengue cases (CD); humidity and dengue cases (HD). Depending on the city, our results show that the climate data can improve or not the forecast. For instance, for Natal, the case D induces a better forecast. For Iquitos, it is better to use all the climate variables. Nonetheless, for Barranquilla, the forecast is better, when we include cases and humidity data. Another important result is that each city has an optimal region based on the training length. For Natal, when we use more than 64% and less than 80% of the time series for training, we obtain results with correlation coefficients (r) among 0.917 and 0.949 and mean absolute errors (MAE) among 57.783 and 71.768 for the D case in forecasting. The optimal range for Iquitos is obtained when 79% up to 88% of the time series is considered for training. For this case, the best case is CD, having a minimum r equal to 0.850 and maximum 0.887, while values of MAE oscillate among 2.780 and 4.156. For Barranquilla, the optimal range occurs between 72% until 82% of length training. In this case, the better approach is HD, where the measures exhibit a minimum r equal to 0.942 and maximum 0.953, while the minimum and maximum MAE vary among 6.085 and 6.669. We show that the forecast of dengue cases is a challenging problem and climate variables do not always help. However, when we include the mentioned climate variables, the most important one is the humidity.

1 Introduction

The number of dengue cases has increased over the past 2 decades. The worldwide reported cases in 2000 were 500,000 but have an incredible increase to 5.2 million in 2019 [1]. Most of these cases were reported by countries in regions of Americas. For several countries belonging to these region, the beginning of 2024 is marked by an exponential increase in the dengue cases [2]. In regions of Americas, 4,565,911 cases were reported in 2023, where 0.17% were severe cases and 0.051% implied in deaths. This problem intensified in 2024, where 673,267 cases were reported already during the first 5 epidemiological weeks. From these amounts, 0.1% were severe and 0.015% of the cases ended in deaths. This way, dengue is a challenging problem for public health and in particular for the dengue outbreak prediction models [3, 4].

[a] e-mail: enriquec@pik-potsdam.de (corresponding author)

The dengue virus is transmitted to humans by bites of infected *Aedes* mosquitoes, mainly by *aegypti* and *albopictus* [5, 6]. There are four known dengue serotypes, which are designated by the acronym followed by the index, as DENV-1, 2, 3, and 4 [7]. Once infected by one serotype, the individual acquires permanent immunity against it and temporary cross-immunity to other ones [8]. Secondary infections increase the probability of dengue hemorrhagic fever [9]. Besides the consequences for human health, this disease has significant impacts on the society and economy [10]. Although the dengue incidence is spatially distributed, higher outbreaks occur in the tropical countries, mainly in South America [11]. Some reasons for that are due to the environmental conditions and climate factors [12], such as temperature [13] and precipitation [14]. The temperature acts directly in the life cycle and reproductive of the mosquitoes [15, 16], while the rain precipitation provides containers for eggs and habitat for the mosquitoes [17, 18]. Moreover, the relationship between rainfall and temperature are essential for regulating the water habits. Therefore, the climate variables play an important role in the seasonality of dengue epidemics [19, 20].

The correlation between climate and dengue cases has been studied by many approaches [20–26], among them Machine learning (ML) techniques [27]. Considering a multi-stage ML approach, Appice et al. investigated the evolution of dengue cases in Mexico [28]. To improve forecasting, they analyzed the temperature influence. Guo et al., by means of five ML algorithms, developed a forecast model for dengue data from China [29]. Additionally, they took into account climate data, such as mean temperature, relative humidity and rainfall. They showed that the support vector regression algorithm is the best one to forecast the outbreaks in China. An ensemble neural network model was proposed to forecast dengue outbreaks based on rainfall data from San Juan, Iquitos and Ahmedabad [30]. The authors used a framework available to forecast long-term cases around 52 weeks. Deep learning techniques were employed by Zhao et al. to study and forecast the dengue cases in Singapore [31]. They also considered the average temperature and rainfall time series. The best performance for 2, 3 and 4 weeks in advanced forecasting was 84.61%. Other algorithms have been employed to forecast dengue cases based on climate data [32–37]. In addition, it is important to mention that a lot of work have been done in Physics using ML techniques in recent years [38]. More specific, in fields such as Social Physics [39], Criminal Networks [40], Chaos in cancer growth [41], and other [42].

Roster et al. used ML based on meteorological variables to forecast dengue cases in Brazil [43]. They considered Random Forest (RF), gradient boosting regression, multi-layer perceptron and support vector regression methods. The dengue cases and meteorological variables were monthly recorded from 2007 to 2019 and from 2005 to 2019, respectively. Training the algorithms, the best performance was obtained by the RF method. With regard to RF, Ong et al. investigated the dengue transmission in Singapore utilizing data from the spread of dengue, such as population, entomological and environmental data [44]. Their results showed that RF has high accuracy to reproduce dengue cases, obtaining a correlation coefficient greater than 0.86. They demonstrated that spatial risk of dengue cases can be modeled by means of RF. For Iquitos (Peru), San Juan (Puerto Rico) and Singapore, Benedum et al. combined weather data with dengue cases to construct a forecast model [45]. Comparing time series, regression and RF methods, they verified that the RF method has 21% and 33% of error less than regression and time series models, respectively, for near short predictions (4 up to 12 weeks). Mussumeci and Coelho also employed the RF method to forecast dengue incidence in 790 cities in Brazil. Concerning the climate data, the authors considered information about incidence cases in social networks.

In this work, we employ the RF method [46] to forecast dengue incidence in three cities localized in South America: Natal (Brazil), Iquitos (Peru) and Barranquilla (Colombia). In our simulations, we use the data delayed by up to one week to forecast the new cases. We also employ three combinations of features: (i) we took into account only dengue cases (D); (ii) we combine dengue and climate data (CD); (iii) we utilize the data of humidity and dengue cases (HD). As an important new finding, we show that, depending on the city and the training length, the results can be improved with a given combination of features. For instance, for Natal when we consider climate variables, the forecasting is not improved. For Iquitos, the strategy ii is better for forecasting, while for Barranquilla the best strategy is iii. Depending on the training length, we find an optimal region for each city where the correlation among real and simulated data increase.

The structure of this work is given by the following order: In Sect. 2, we describe the data acquisition processing and the RF method. Section 3 is dedicated to exhibit and extract information of each time series. Forecasting results are discussed in Sect. 4. Finally, our conclusions are drawn in Sect. 5.

2 Methods

2.1 Data acquisition

In this work, we consider weekly dengue cases and average week climate variables from three localities: (i) Natal (Brazil, elevation 30 m, latitude -5.81 and longitude -35.25) from the 16th week of 2016 until the 52th week of 2019, with time series totalling a length equal to 193 weeks; (ii) Iquitos (Peru, elevation 106 m, latitude -3.74 and

longitude -73.25) from the 28th week of 2001 until the 52th week of 2012, whose time series length is equal to 597 weeks; (iii) Barranquilla (Colombia, elevation 18 m, latitude 10.96 and longitude -74.79) from the 2nd week of 2011 until the 47th week of 2016, with a time series length of 307 weeks. For Natal, we extract the dengue cases from Sanchez-Gendriz et al. [47] and the climate data (precipitation, relative humidity and air temperature) from National Institute of Meteorology [48]. For Iquitos, we obtain both data from the Dengue Forecasting Project Data Repository [49]. The climate variables for Iquitos are: minimum and average temperature, relative and absolute humidity and precipitation. Dengue cases from Barranquilla are available on Sivigila Portal [50] and meteorological (maxima temperature, relative humidity and precipitation) in Ref. [51].

The data and codes employed in this research are available on GitHub [52].

2.2 Data processing and statistical analysis

When there are many outliers in the data set, a standardization technique helps in decreasing the error in the results. In this work, we use the Robust Scaler method, which works by subtracting the median ($\mathrm{med}(X)$) of the data (X) and scaling in the interval between the 1st (Q_1) and the 3rd (Q_3) quartiles. The Robust Scaler equation is given by

$$\mathrm{RS}(x_i) = \frac{x_i - \mathrm{med}(X)}{Q_3 - Q_1}. \tag{1}$$

In this procedure, the median and interquartile range ($Q_3 - Q_1$) are stored and used on future data as the transformation applied on the forecasting.

One important characteristic of time series is to know if it is stationary or non-stationary. There are some techniques that are employed to answer this question. In this work, we consider the augmented Dickey–Fuller test (ADF), which belongs to the unit root test. A unit root test verifies if the time series is non-stationary (null hypothesis) or not (alternative hypothesis). The ADF test can be described by

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t, \tag{2}$$

where α is a constant, β is the time coefficient, γ is the coefficient that present the root (is the focus of test), p is the lag order of the first differences in autoregressive process, and ϵ is a noise term. The parameter that we focus is γ. If $\gamma = 0$, or positive, then we assume the null hypothesis, i.e., a non-stationary time series. On the other hand, if γ is negative we assume the stationarity. In addition to this analysis, we also test the null hypothesis by computing the p-value. If p-value < 0.05 we assume that the time series is stationary. We conducted the ADF test by the URCA package in R [53].
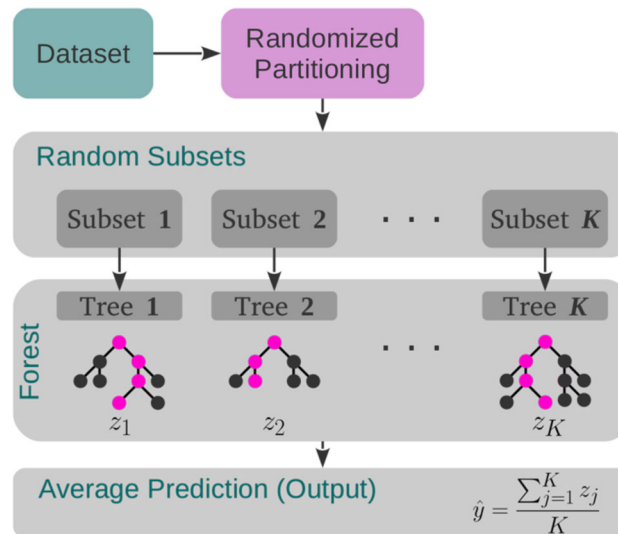
2.3 Random forest

We adopt here the random forest (RF) algorithm [54], which is a supervised ML algorithm based on the ensemble learning method. It is formed by many decision trees. The ensemble learning techniques are more accurate to made predictions than individual models. It occurs due to the fact that the ensemble techniques combine many basic ML algorithms and their predictions. Given an initial data set, the model splits it in K random subsets, generating different subsets from the original data set. These final subsets are namely terminal or leaf nodes and the intermediate subsets are called internal nodes. The prediction of the results in each terminal node is made by the average outcome of the training data. In this way, the RF algorithm generates predictions or rankings from a set of multiple decision trees, as schematically represented in Fig. 1. One advantage of using RF is its ability to handle complex data sets and mitigate overfitting. One criterion to stop the iteration is based on the node size, which specifies the minimum number of observations in a given terminal node [55, 56]. The capacity of RF to handle with complex data sets without leading to overfitting is related to the choice of the hyperparameters (criteria to choose partitions, number of trees, maximum features, and depth) of the model. To identify the optimal parameters, we use the GridSearchCV package from the sklearn library [57]. This process is stopped when the smallest error is obtained as a function of the number of trees [58].

By combining the outputs of the trees, the algorithm provides a consolidated and more accurate result than the result from basic ML algorithms.

The choice of partitions in each decision tree (Fig. 1) is made by considering the minimum mean squared error (MSE)

$$\mathrm{MSE}_{\min} = \min\left\{ \sum_{i \in S_k} (y_i - \widehat{y}_i)^2 \right\}, \tag{3}$$

Fig. 1 Schematic representation of the random forest algorithm. The initial data set is randomly partitioned by the algorithm creating K subsets. These subsets are processed by K decision trees, resulting in z_K. Thereafter, an average prediction overall z_K is taken, which is the output from the model. K random bootstrap samples are extracted from the data and an unpruned decision tree is fitted for each bootstrap sample. At each node, a small subset of the covariance is randomly chosen to optimize the split. The forecasting is obtained by averaging the prediction of all trees

where $\widehat{y}$ is the average value of the partition and y_i is the value of each data point within that partition. Then, it is important to use the prune method as a regularization. This method makes a balance between the lowest MSE value and its depth according to

$$\mathrm{MSE}_{\mathrm{min},\,\alpha} = \min_{\alpha}\left\{\sum_{i \in S_k}(y_i - \widehat{y}_i)^2 + \alpha T\right\}, \tag{4}$$

where α is a tuning parameter that is determined by cross validation and T is the number of terminal nodes [59]. Our simulations are implemented in Python and we use sklearn libraries for statistical analyses [57, 60].

For the training process, we use 70% of the time series length and for testing 30%. Then, based in the inputs y_i (dengue cases), p_i (precipitation), h_i (humidity), T_i (temperature), in the i-th week, our algorithm forecasts the dengue cases in y_{i+1} by searching a function f such that

$$y_{i+1} = f(y_i,\, p_i,\, h_i,\, T_i). \tag{5}$$

In this work, we use three combinations of inputs: climate data and dengue cases (CD); dengue cases (D); humidity and dengue cases (HD).

To obtain the best hyperparameters (criteria to choose partitions, number of trees, maximum features, and depth), we use the grid method search [61]. For Natal, we split the time series into 144 weeks for training and 49 for testing. In this case, the best hyperparameters are, respectively: Friedman MSE [59], 1000, 0.5, and 100. For Iquitos, the data are split into 333 weeks for training and 263 for testing. In this case, the hyperparameters are Absolute Error, 200, 0.6, and 50. Finally, for Barranquilla, the time series is split into 233 weeks for training and 74 weeks for forecasting. The hyperparameters in this case are: Poisson, 1500, 0.2, and 200. We also consider the bootstrapping active.

2.4 Error analysis

In the forecasting range, we compute the error among simulated and real points by considering the absolute error (e), mean absolute error (MAE), and the error (ΔE). Given the real time series denoted by y_i and the simulated points x_i, the absolute error in the i-week is defined as

$$e_i = |y_i - x_i|, \tag{6}$$

Table 1 Test of stationarity using ADF test for the Natal features

Feature	p-value	γ
Cases	8×10^{-2}	-0.053 ± 0.024
Precipitation	2×10^{-16}	-0.803 ± 0.070
Rel. humidity	10^{-10}	-0.421 ± 0.058
Min. temp.	8×10^{-4}	-0.150 ± 0.040
Max. temp.	4×10^{-5}	-0.202 ± 0.044
Avg. temp.	9×10^{-4}	-0.143 ± 0.037

the notation e_{D}, e_{CD} and e_{HD} means the absolute error computed using x_i from strategy D, CD, and HD, respectively. Where the index i is omitted by economy of notation. Another measure that we use is

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|, \tag{7}$$

where n is the number of weeks.

Another important question to study is whether our model leads to overestimating or underestimating cases. To extract this information, we compute

$$\Delta E_i = y_i - x_i, \tag{8}$$

then, when $\Delta E_i < 0$ the model overestimating, and $\Delta E_i > 0$ the forecasting is underestimating.

3 Time series

3.1 Natal (Brazil)

Time series for Natal are displayed in Fig. 2, where the panel (a) shows the dengue cases ($\times 10^2$), (b) precipitation (mm) and (c) relative humidity (%). The panel (d) exhibits the minimum (blue line), maximum (red line) and average temperature (black line) in °C. The bottom x-axis displays the year and its relative epidemiological week, while the upper x-axis shows the number of week for the whole time series. We compute the power spectra density using the package PDSR implemented in R [62] for all the variables. Our results show that the period is approximately 1 year for all variables. Specifically, for the dengue cases we obtain a period equal to 53 ± 2 weeks. Before employing the ML algorithm, we check the stationary of the data by means of the ADF method. The result of the test is displayed in Table 1. As the p-value for the dengue cases is greater than 0.05 and γ is close to zero, we conclude that this time series is non-stationary. Figure 2a exhibits the data during the four analyzed years. From the 16th week of 2016 until the end of this year, 3422 cases were reported. In 2017, 2018 and 2019, there were 4747, 15,178, and 17,197 cases reported, respectively. In 2018 and 2019, there were more cases than expected, then in these years outbreaks occurred. Considering the time range of the first outbreak from the 10th until the 37th week of 2018 (where more than 200 cases occur every week), there were 12,442 reported cases. Using the same criteria, the second outbreak exists between the 12th until the 38th week of 2019, where 14,687 cases were reported. Just these two outbreaks were responsible for 27,129 infections.

3.2 Iquitos (Peru)

The time series for the data from Iquitos are displayed in Fig. 3, where the panel (a) is for dengue cases ($\times 10$), (b) is for precipitation (mm), (c) is for absolute humidity (%), (d) for relative humidity (%) and (e) is for minimum (blue line) and average temperature °C (black line). The considered range for these data varies from the 28th week of 2001 until the 52th week of 2012, totalling 597 weeks (marked in the top x-axis). By the power spectra density analyze, our results shows a period of 1 year for all the variables. More precisely, the cases has a period equal to 55 ± 3 weeks. The ADF test shows that the time series related to Iquitos are stationary, as observed in Table 2. The associated p-values in Table 2. In terms of reported cases, Iquitos received 274, 490, 171, 715, 451, 256, 562, 694, 296, 585, 95, and 501 reports during 2001–2012, respectively. Along this time series, we observe the presence of 9 outbreaks along these 12 years. Two of them call attention due to the high amplitude. The first one

Fig. 2 **a** Dengue cases ($\times 10^2$), **b** precipitation (mm), **c** relative humidity (%) and **d** air temperature (°C) for Natal. The time series is from the 16th week of 2016 until the 52th week of 2019 (in bottom x-axis), resulting in 193 weeks (top x-axis
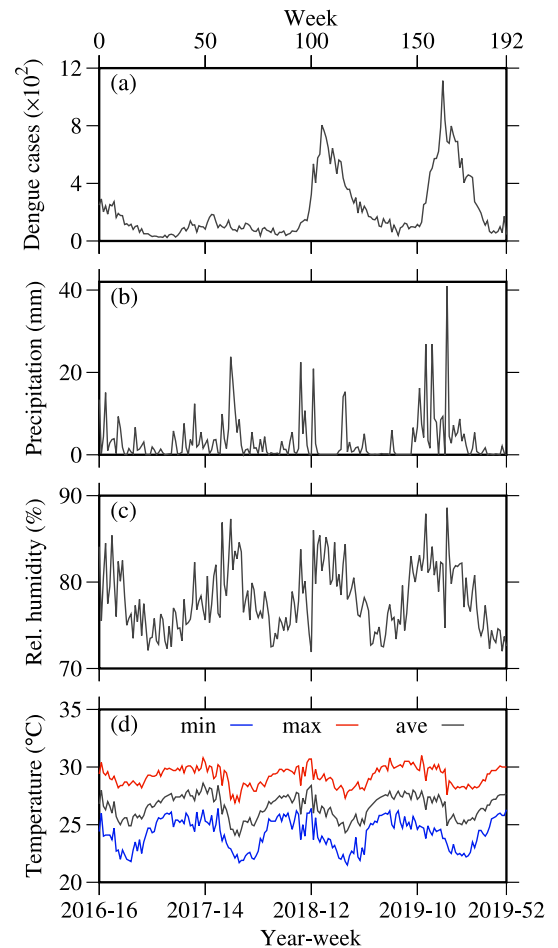
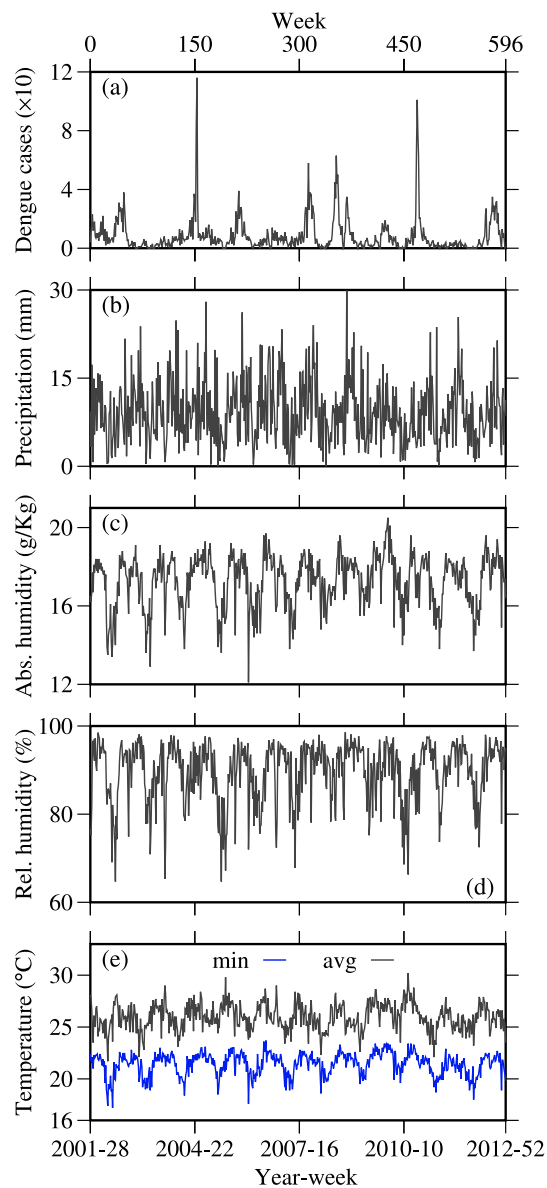**Table 2** Test of stationarity using ADF test for the Iquitos features

Feature	p-value	γ
Cases	10^{-14}	-0.221 ± 0.024
Precipitation	2×10^{-15}	-0.855 ± 0.040
Spec. humidity	2×10^{-16}	-0.414 ± 0.033
Rel. humidity	2×10^{-16}	-0.322 ± 0.030
Min. temp.	2×10^{-16}	-0.300 ± 0.030
Avg. temp.	2×10^{-16}	-0.432 ± 0.033

occurs between the 22th and the 27th weeks of 2004 and the second between the 25th and the 35th weeks of 2010, occurring a total of infection equal to 265 and 413, respectively.

3.3 Barranquilla (Colombia)

Figure 4 displays the time series for Barranquilla. The panels (a), (b), (c), and (d) exhibit the time evolution of the dengue cases ($\times 10^2$), maximum air temperature °C, precipitation (mm) and relative humidity (%), respectively. The data for Barranquilla start in the first week of 2011 and end in the 47th week of 2016 (bottom x-axis), totalling 307 weeks (top x-axis). From the spectral analyzes, the period associated with the climate variables is 1 year. The period for dengue cases in Barranquilla correspond to 52 ± 4 weeks. The time series from Barranquilla are stationary, as observed in Table 3. If we consider an outbreak when more than 50 cases are reported, we see 3 outbreaks in Fig. 4a. The first one starts in the 8th week of 2013 and ends in the 52th week of the same year, having 2425 reported cases. The second one starts in the 39th and finishes in the 53th week of 2014 where 1210 cases were reported. The last one occurs from the 43th week of 2015 and finishes in the 3rd week of 2016, with

Fig. 3 Time series from Iquitos, where the panel **a** shows the dengue cases (×10), **b** precipitation (mm), **c** absolute humidity (g/kg), **d** relative humidity (%) and **e** temperature (°C). The data are from the 28th week of 2001 until the 52th week of 2012 (bottom x-axis), totalling 597 weeks (top x-axis)

**Table 3** Test of stationarity using ADF test for the Barranquilla features

Feature	p-value	γ
Cases	10^{-4}	-0.116 ± 0.027
Max. temp.	3×10^{-11}	-0.293 ± 0.040
Precipitation	10^{-16}	-0.696 ± 0.054
Rel. humidity	2×10^{-10}	-0.271 ± 0.040

979 cases. Due to the huge outbreaks in 2013 and 2014, they are the years with more reported cases. For instance, 655, 946, 1364 and 617 cases were reported in 2011, 2012, 2015, and 2016. However, the data show 2748 and 2737 cases in 2013 and 2014, respectively.

Fig. 4 Time series from Barranquilla. The panel **a** displays the dengue cases ($\times 10^2$), **b** precipitation (mm), **c** relative humidity (%) and **d** temperature (°C). The data were collected from the 2nd week of 2011 up to the 47th week of 2016 (bottom x-axis), resulting in 307 weeks (top x-axis)
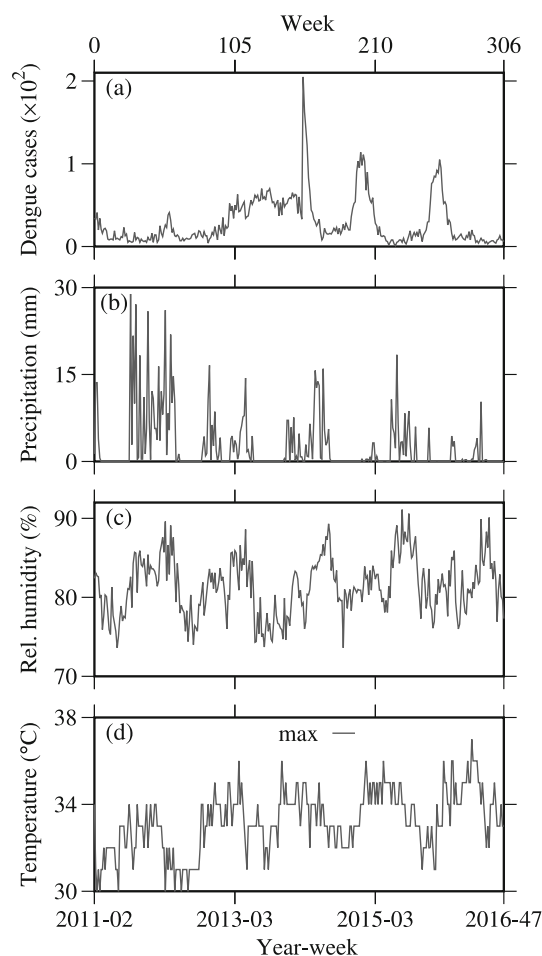

4 ML-based forecasting

4.1 Natal forecasting

First, we employ the algorithm in Natal with training a length equal to 144 weeks and the 49 remaining weeks. In our algorithm, we consider the number of trees, maximum features and depth equals to 1000, 0.5 and 100, respectively. The criteria to choose partitions is Friedman MSE [59]. For CD, we consider previous cases, minimum, maximum and average temperature, humidity and precipitation. For each feature, the algorithm gives the following importance: 0.67, 0.04, 0.03, 0.03, 0.12, and 0.08. The dengue cases and humidity are more often used by the algorithm. Figure 5a displays the dengue cases (black points) and our simulated results (colored lines). The training region shows the red (CD approach) and the dotted blue (D approach) line, while the test range (gray background) exhibits the magenta (CD) and green (D) lines. As expected, in the train region, we observe a good accordance among the points generated by the ML algorithm and the real data. In the test region, magnified in the panel (b), the error increases when compared with the training. In the panel (b), the new curves exhibit the absolute error (e) among the real and simulated points. The light magenta color, denoted by e_{CD}, is the absolute error that is associated with the CD approach, while the light green (e_D) with D approach. Our results show that the error increases in the peak region, in both approaches. The mean absolute error (MAE) for CD in the test range is 97.92 and the correlation coefficient (r) is 0.90, while for D approach these values change to 62.27 and 0.94, respectively. Considering that the peak starts in the 156th week and ends in the 173th week, the MAE before the peak is 49.24 and 42.08 for the CD and D approach. During the peak, the MAE is 166.67 and 88.34 for CD and D, respectively. After the peak the MAE is 35.44 and 52.22 for CD and D. During and before the peak, D performs better than the CD approach. This scenario changes after the peak. During the test region, the maximum value of reported cases is 1113 and the MAE in peak range represents approximately 14.9% and 7.9% of this value. Outside this range, the MAE is around 4%. The result for ΔE_i in the testing range, is displayed in Fig. 5c by the blue line for CD and the black line for D. Before the peak, we have mostly underestimating values.

Fig. 5 Training and forecasting results using CD and D strategies for Natal dengue cases. **a** Dengue cases (black points) and ML forecast results for Natal. The red and dotted blue lines show the training range for CD and D approach, respectively. The test range is highlighted by the gray background, where the magenta and green lines display the results by considering the CD and D approaches, respectively. **b** Magnification of testing region. The light magenta and blue curves display the absolute error e among real data and CD and D approach, respectively. **c** Error among real and simulated data
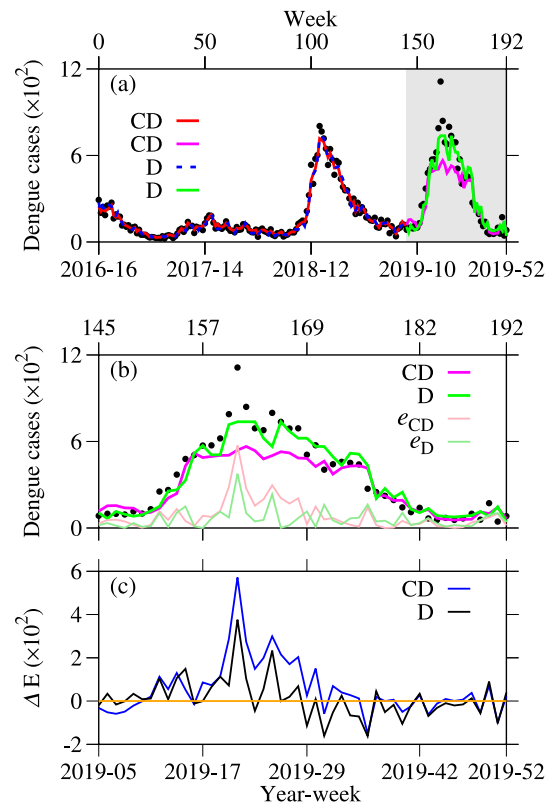
After the peak, we identify a mix. This overestimation occurs due to the fact that the algorithm learns this huge peak and this information remains in its memory.

Figure 6 displays the comparison between each approach, i.e., CD (blue line), D (dotted black line) and HD (dotted red line), as a function of the training length. The panel (a) exhibits the MAE and the panel (b) shows the r computed in the remaining test region. The main results are: when we train our algorithm for a few weeks, i.e., less than 55% of the time series, the RF produces considerable error in the forecast. Increasing the training length, the RF starts to perform better in the forecast. If we consider more than 90% of our time series for training, the error increases and the correlation decreases. After 90% of training, just a few weeks remain for forecast, then we do not have enough weeks to obtain a reasonable statistic. Utilizing between 60% and 89% of the time series for training, we observe that the MAE decreases and r increases. In this range, a better performance is obtained for the D approach. For a training length inferior to 55%, the CD and HD perform better than D approach.

Fig. 6 Performance estimation of CD, D and HD approaches as a function of the training length for Natal data. **a** Mean absolute error (MAE) and **b** correlation coefficient (r) as function of the length training. The blue line displays the results for CD approach, while the dotted black and red lines for D and HD, respectively
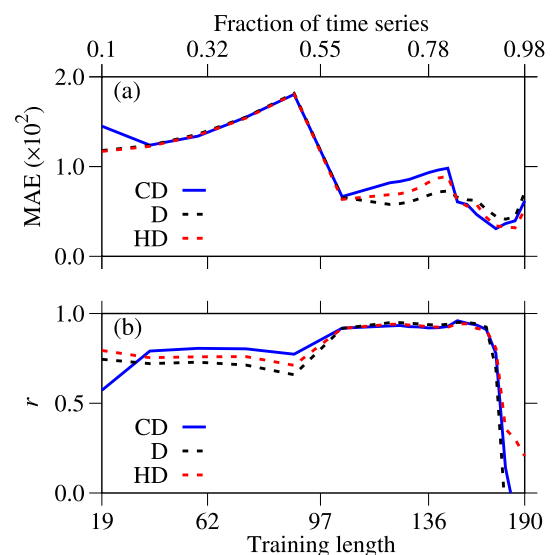
4.2 Iquitos forecasting

The second analyzed city is Iquitos. There we consider 333 weeks of training and 263 of forecasting. We choose this training length (56% of the whole time series) due the fact that those 333 weeks separated two, apparently, symmetric outbreaks. The result is exhibited in Fig. 7a, where the red and dotted blue lines are related to training and the magenta and green lines to the test region for the CD and D approach, respectively. The black points exhibit the dengue cases ($\times 10$). The hyperparameters are the number of trees equal to 200, maximum features equal to 0.6, depth equal to 50 and criteria to choose partitions equal to the Absolute Error. In the CD approach, we consider the following features: Cases, minimum and average temperature, relative and absolute humidity, and precipitation. The algorithm weights these features as 0.39, 0.11, 0.10, 0.13, 0.13 and 0.13, respectively, in the learning process. For this, humidity and precipitation are mostly used. In addition, we consider a very long horizon forecast (up to 263 weeks). We obtain one MAE = 4.42 and $r = 0.83$ in the test range for the CD situation. For D approach, these values change to 4.02 and 0.81, showing that CD performs better than D in this case. The first outbreak in the testing range has a symmetric series of outbreaks. Due to this reason, we choose this training length. Doing this consideration, we observe that RF performs a good forecast, as displayed in the magnification in Fig. 7b. As in the previous subsection, the algorithm also fails in predicting the highest peaks. The better performance in this situations seems to be in the D approach, however, this is not true. By looking for the MAE along the time series, as exhibited in Fig. 7c, we verify that the error associated with the CD approach (light magenta line, denoted by e_{CD}) is less in the outbreaks when compared with the errors associated with the D situation (light green line, denoted by e_{D}). Also, after the huge peak localized at the 467th week, the D approach exhibits an overestimation of cases (panel (e)) higher than the CD approach (panel (d)).

MAE and r can be improved if we train our algorithm for longer time (Fig. 8). If we take less than 55% of the time series to train the algorithm, MAE increases and r decreases. In addition, for certain values, CD (blue line) performs worse than D (black dotted line) or HD (red dotted line) approaches. For values after week 331 (55% of time series) of training, the results start to improve, in the sense of obtaining a minimum MAE and maximum r, until a threshold value that occurs when we use 90% of the time series for training. At this point, MAE and r are respectively equal to 3.57 and 0.88 for CD, 4.06 and 0.77 for D, and 3.53 and 0.85 for HD. We see that CD and HD perform better than D. After 90%, MAE increases and r decreases. Besides, there are enough weeks after this limit (less than 60). In this given range, the algorithm does not perform very well in forecasting. The range captures just the last outbreak of the time series, where the points oscillate very irregularly.

Fig. 7 Training and forecasting results using CD and D strategies for Iquitos dengue cases. **a** Iquitos dengue cases (black points) and ML forecast results. The red and dotted blue lines show the training range for the CD and D approaches, respectively. The test range is highlighted by the gray background, where the magenta and green lines display the results by considering the CD and D approaches, respectively. **b** Magnification of testing region. **c** MAE among real and simulated data, where the light magenta line is associated with CD and the light green with D. **d** Absolute error for the CD approach. **e** Absolute error for the D approach
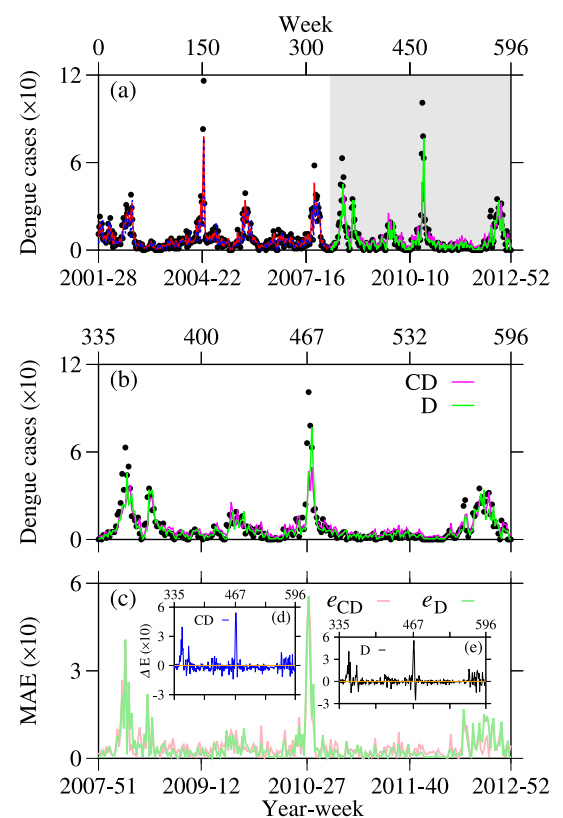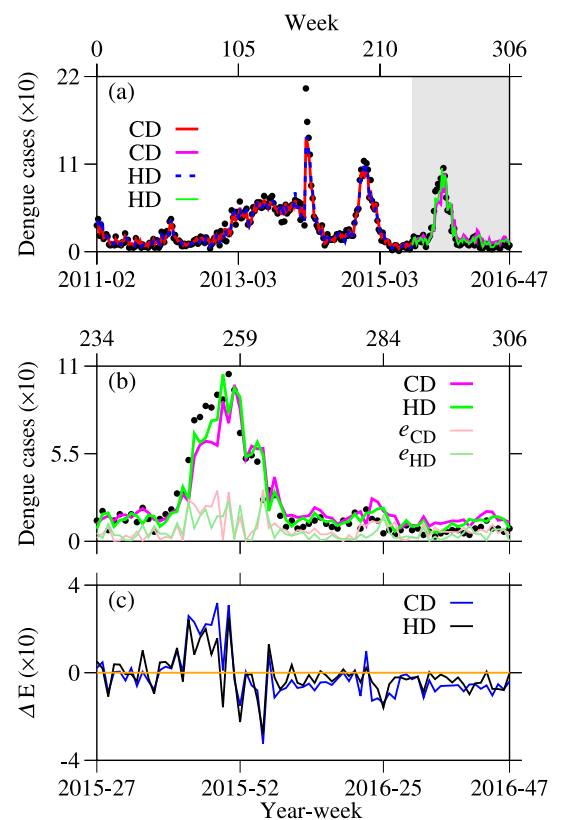
Fig. 8 Performance
estimation of CD, D and
HD approaches as a
function of the training
length for Iquitos data.
a Mean absolute error
(MAE) and **b** correlation
coefficient (r) as a function
of length training. The blue
line displays the results for
CD approach, while the
dotted black and red lines
for D and HD, respectively
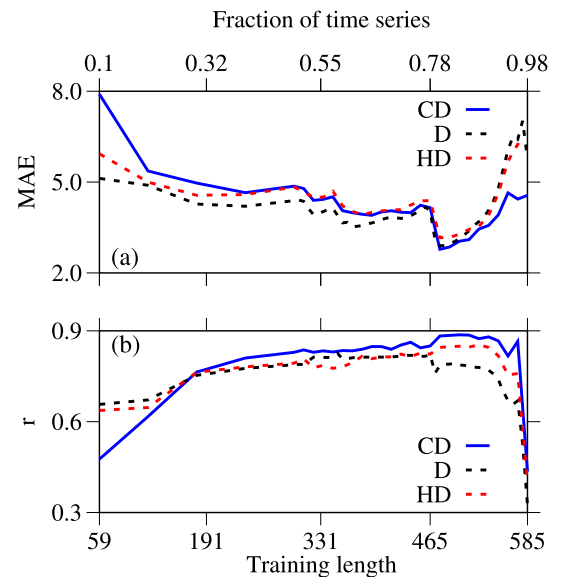
4.3 Barranquilla forecasting

The last analyzed city is Barranquilla. For this city, our simulations show better results using the HD approach than the D approach. Figure 9a displays the real data for dengue cases ($\times 10$) by the black points, red and blue lines the training range, and magenta and green the testing curves for CD and HD approaches, respectively. In our simulation, we take 233 weeks for training, corresponding to 76% of the time series length and remaining 74 weeks of forecasting. We use this training length to observe the forecast of the last outbreak. For the hyperparameters the number of trees, maximum features, depth, and criteria to choose partitions, we utilize 1500, 0.2, 200, and Poisson. In the CD approach, the ML uses 0.65, 0.10, 0.16, and 0.10 of the features cases, maximum temperature, relative

Fig. 9 Training and
forecasting results using CD
and HD strategies for
Barranquilla dengue cases.
a Barranquilla dengue cases
(black points) and ML
forecast results. The red
and dotted blue lines show
the training range for CD
and HD approaches,
respectively. The test range
is highlighted by the gray
background, where the
magenta and green lines
display the results by
considering the CD and HD
approaches, respectively.
b Magnification of testing
region. **c** Error among real
and simulated data

Fig. 10 Performance estimation of CD, D, and HD approaches as a function of the training length for Barranquilla data. **a** Mean absolute error (MAE) and **b** correlation coefficient (r) as a function of length training. The blue line displays the results for the CD approach, while the dotted black and red lines for D and HD approaches, respectively
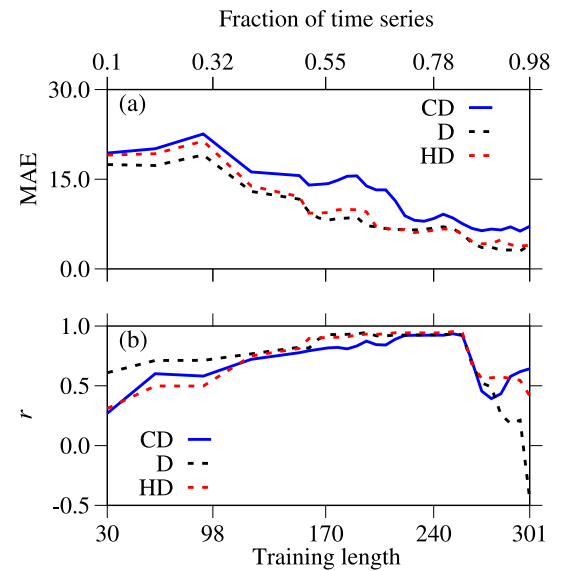
humidity, and precipitation, respectively. In the HD features, the algorithm provides 0.82 and 0.17 of importance to cases and relative humidity, respectively. For the first approach, we obtain 7.81 for MAE and 0.92 for r, while for HD, the values are 6.22 and 0.94. The testing range is amplified in Fig. 9b, where the light magenta and green curves show e_{CD} and e_D. It is important to note that the fitting of the peak performs better than in the Natal case. Considering the peak region defined between the 250th until the 262th week, the CD approach returns one MAE equal to 15.90 and HD equal to 13.16. As also observed in Natal, the approach without all considered climate variables performs better in the peak range. From the point where we started the training until 249 the MAE for CD and HD are, respectively, 3.63 and 5.00. After 262 weeks, these values are 7.09 and 4.69. The CD approach performs better before the peak. Another characteristic that emerges after 262 weeks is that the ML overestimates the forecast values, as observed by the results in Fig. 9c.

A comparison among the approaches CD (blue line), D (black dotted line), and HD (red dotted line) is displayed in Fig. 10, where panel (a) exhibits MAE and panel (b) shows r as a function of training length. As observed in the previous results, the algorithm performs better in the testing region, when we consider more than 55% of the time series as training. However, in this case, if we use more than 85% of the time series as training, the error increases and the correlation decreases. It is important to note that if we take the approach D and more than 96% of the time series as training, the correlation becomes negative. The algorithm forecast increases in the cases when the data show a decay. Comparing the methods, D performs better in terms of MAE for a wider training length when compared with the other two methods. However, we look for the lower MAE and maximum r simultaneously. Therefore, in the range 55% and 85% the best strategy is use the HD approach.

5 Conclusion

In this work, we employ random forest (RF) machine learning (ML) technique to forecast dengue infections based on previous cases and meteorological variables. We test our approach for data from Natal (Brazil), Iquitos (Peru), and Barranquilla (Colombia). In our simulations, we use three approaches: where we consider only dengue cases (D); a combination of climate variables and dengue cases (CD); the combination of humidity and dengue cases (HD). We use the features from the i-week to forecast the cases in $(i+1)$-week. We also test different delays in the three cities; however, we do not obtain improvement. For each city, we use a set of climate variables. For Natal, we utilize the average, minimum, maximum temperature, precipitation and relative humidity. We find that the algorithm uses more humidity and the dengue cases in the forecasting process. Nevertheless, for Natal, we get an improvement including the dengue cases in the prediction only. For Iquitos, we consider the relative humidity, minimum and average temperature, specific humidity and precipitation. In this case, the humidity and precipitation are used with more weight by the ML technique. The best forecasting performance for Iquitos occurs, when climate and dengue cases are included. For Barranquilla, we utilize the maximum air temperature, relative humidity and precipitation. Then, the most important variable considered by the ML is the relative humidity. For the data from Barranquilla, our results show a better performance, when there is the combination of dengue and humidity data.

A common characteristic that emerges is that the forecast generated by the RF algorithm exhibits a higher error in the peak regions. Moreover, exploring the effects of training length, we find that the algorithm performs better when we use more than 55% of the time series and less than 90%. The optimal region for Natal occurs when we include more than 64% and less than 80% of the time series for training. In this range, the error decreases and the correlation increases. The best performance is generated by the D method. For this approach, we verify correlations varying between 0.917 and 0.949, and MAE between 57.783 and 71.768. The optimal range for Iquitos occurs considering between 79% and 88% of the time series as training length. The method that performs better is CD, having a minimum r equal to 0.850 and maximum equal to 0.887 while MAE oscillates between 2.780 and 4.156. For Barranquilla, the optimal range occurs between 72 and 82% of length training. The method that performs better in this range is HD with minimum r equal to 0.942 and maximum equal to 0.953 while the minimum and maximum MAE are equal to 6.085 and 6.669, respectively. The ML technique reproduces many outbreaks and after that there is no outbreak. Then, the algorithm overestimated some cases, as we observed in Barranquilla. Another situation is that there are only few points in order to do statistical analysis, as in Natal. For Iquitos, the high values of training length select a very noisy region in the time series. The region is characterized by one outbreak and many fluctuations that are not predicted by RF.

Besides, the dengue vector needs proper climate conditions. Our work exhibits that there is not a preferable climate variable in the forecasting process using the RF method. On the other hand, climate variables can increase the error and decrease the correlation in some situations. In addition, also as a novelty, we find that the humidity can be more relevant than other climate variables in the forecasting process. Our results were also tested using the classic algorithms CNN 1D, LSTM, ARIMA, and SARIMA regression and also we normalize the humidity data removing the annual component. However, the RF algorithm displays a better result for our goal and the modification in the humidity did not improve our results. In this way, our work shows that dengue forecasting is a challenging problem. The climate variables need to be carefully selected in order to improve the prediction or not. Moreover, we show that under specific conditions it is better use just the dengue cases (D) as input.

Our work shows that the RF method is useful for the forecasting of new dengue cases based on meteorological characteristics. This methodology can be employed by public health organization to forecast and study control measures, because RF is fast, efficient, robust, and exhibits a high correlation and low mean absolute error in the forecasting. For future works, we plan to explore other characteristics, such as mosquito eggs, and contamination rates to better predict new cases.

Data availability The datasets generated during and/or analyzed during the current study are available on GitHub [52] and also can be requested from the corresponding author.

References

1. Dengue emergency in the Americas: time for a new continental eradication plan. The Lancet Regional Health—Americas. https://doi.org/10.1016/j.lana.2023.100539. Accessed 08 Mar 2024
2. Epidemiological Alert—Increase in dengue cases in the Region of the Americas—16 February 2024. Pan American Health Organization. https://www.paho.org/en/documents/epidemiological-alert-increase-dengue-cases-region-americas-16-february-2024. Accessed 08 Mar 2024
3. S. Bhatt, P.W. Gething, O.J. Brady, J.P. Messina, A.W. Farlow, C.L. Moyes, J.M. Drake, J.S. Brownstein, A.G. Hoen, O. Sankoh, M.F. Myers, D.B. George, T. Jaenisch, G.R.W. Wint, C.P. Simmons, T.W. Scott, J.J. Farrar, S.I. Hay, Nature **496**, 504–507 (2013)
4. World Health Organization, *Dengue* (WHO, Geneva, 2023)

5. L. Rosen, D.A. Shroyer, R.B. Tesh, J.E. Freier, J.C. Lien, Am. J. Trop. Med. Hyg. **32**, 1108–1119 (1983)
6. World Health Organization, *Dengue and Dengue Hemorrhagic Fever* (WHO, Geneva, 2023)
7. J.P. Messina, O.J. Brady, T.W. Scott, C. Zou, D.M. Pigott, K.A. Duda, S. Bhatt, L. Katzelnick, R.E. Howes, K.E. Battle, C.P. Simmons, S.I. Hay, Trends Microbiol. **22**, 138–146 (2014)
8. C.P. Simmons, J.J. Farrar, N. van Vinh Chau, B. Wills, N. Engl. J. Med. **366**, 1423–1432 (2012)
9. S.B. Halstead, Adv. Virus Res. **60**, 421–467 (2003)
10. D.S. Shepard, E.A. Undurraga, Y.A. Halasa, J.D. Stanaway, Lancet Infect. Dis. **16**, 935–941 (2016)
11. European Centre for Disease Prevention and Control. Dengue world-wide overview. Available on: https://www.ecdc.eu ropa.eu/en/dengue-monthly. Accessed 13 June 2024
12. C.W. Morin, A.C. Comrie, K. Ernst, Environ. Health Perspect. **121**, 11–12 (2013)
13. D.M. Watts, D.S. Burke, B.A. Harrison, R.E. Whitmire, A. Nisalak, Am. J. Trop. Med. Hyg. **36**, 143–152 (1987)
14. J.M. Reinhold, C.R. Lazzari, C. Lahondere, Insects **9**, 158 (2018)
15. A. de Garín, R.A. Bejarán, A.E. Carbajo, S. de Casas, N.J. Schweigmann, Int. J. Biomed. **44**, 148–156 (2000)
16. L. Eisen, A.J. Monaghan, S. Lozano-Fuentes, D.F. Steinhoff, M.H. Hayden, P.E. Bieringer, J. Med. Entomol. **51**, 496–516 (2014)
17. B.W. Alto, S.A. Juliano, J. Med. Entomol. **38**, 646–656 (2001)
18. J. Lega, H.E. Brown, B. Barrera, J. Med. Entomol. **54**, 1375–1384 (2017)
19. M.U.G. Kraemer, R.C. Reiner, O.J. Brady, J.P. Messina, M. Gilbert, D.M. Pigott, D. Yi, K. Johnson, L. Earl, L.B. Marczak, S. Shirude, N.D. Weaver, D. Bisanzio, T.A. Perkins, S. Lai, X. Lu, P. Jones, G.E. Coelho, R.G. Carvalho, W.V. Bortel, C. Marsboom, G. Hendrick, F. Schaffner, C.G. Moore, H.H. Nax, L. Bengtsson, E. Wetter, A.J. Tatem, J.S. Brownstein, D.L. Smith, L. Lambrechts, S. Cauchemez, C. Linard, N.R. Faria, O.G. Pybus, T.W. Scott, Q. Liu, H. Yu, G.R.W. Wint, S.I. Hay, Nat. Microbiol. **4**, 854–863 (2019)
20. L.L. Xavier, N.A. Honório, J.F.M. Pessanha, P.C. Peiter, PLoS One **16**, e0251403 (2021)
21. R. Lowe, C.A. Coelho, C. Barcellos, M.S. Carvalho, R.D.C. Catão, G.E. Coelho, W.M. Ramalho, T.C. Bailey, D.B. Stephenson, X. Rodó, eLife **5**, e11285 (2016)
22. M. Oki, T. Yamamoto, PLoS One **7**, e48258 (2012)
23. B. Paul, W.L. Tham, Health **7**, 672–678 (2015)
24. M.A. Johansson, F. Dominici, G.E. Glass, PLoS Negl. Trop. Dis. **32**, e382 (2009)
25. M.A. Johansson, N.H. Reich, A. Hota, Sci. Rep. **6**, 33707 (2016)
26. H.-Y. Yuan, J. Liang, P.-S. Lin, K. Sucipto, M.M. Tsegaye, T.-H. Wen, S. Pfeiffer, D. Pfeiffer, Sci. Rep. **10**, 4297 (2020)
27. T.M. Carvajal, K.M. Viacrusis, L.F.T. Hernandez, H.T. Ho, D.M. Amalin, K. Watanabe, BMC Infect. Dis. **18**, 183 (2018)
28. A. Appice, Y.R. Gel, I. Iliev, V. Lyubchich, D. Malerba, IEEE Access **8**, 52713–52725 (2020)
29. P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang, W. Ma, PLoS Negl. Trop. Dis. **11**, e0005973 (2017)
30. M. Panja, T. Chakraborty, S.S. Nadim, I. Ghosh, U. Kumar, N. Liu, Chaos Solitons Fractals **167**, 113124 (2023)
31. X. Zhao, K. Li, C.K.E. Ang, K.H. Cheong, Chaos Solitons Fractals **168**, 113170 (2023)
32. M. Cabrera, J. Leake, J. Naranjo-Torres, N. Valero, J.C. Cabrera, A.J. Rodríguez-Morales, Trop. Med. Infect. Dis. **7**, 322 (2022)
33. M.E. Francisco, T.M. Carvajal, M. Ryo, K. Nukazawa, D.M. Amalin, K. Watanabe, Sci. Total Environ. **792**, 148406 (2021)
34. N.A.M. Salim, Y.B. Wah, C. Reeves, M. Smith, W.F.W. Yaacob, R.N. Mudin, R. Dapari, N.N.F.F. Sapri, U. Haque, Sci. Rep. **11**, 939 (2021)
35. M.S. Rahman, C. Pientong, S. Zafar, T. Ekalak-Sananan, R.E. Paul, U. Haque, J. Rocklov, H.J. Overgaard, One Health **13**, 100358 (2021)
36. N. Ochida, M. Mangeas, M. Dupont-Rouzeyrol, C. Dutheil, C. Forfait, A. Peltier, E. Descloux, C. Menkes, Environ. Health **21**, 20 (2022)
37. D.K. Ming, N.M. Tuan, B. Hernandez, S. Sangkaew, N.L. Vuong, H.Q. Chanh, N.V.V. Chau, C.P. Simmons, B. Wills, P. Georgiou, A.H. Holmes, S. Yacoub, Front. Digit. Health **4**, 849641 (2022)
38. G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Nat. Rev. Phys. **3**, 422–440 (2021)
39. M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romic, Z. Wang, S. Geck, T. Lipic, B. Podobnik, L. Wang, W. Luo, T. Klanjscek, J. Fan, S. Boccaletti, M. Perc, Phys. Rep. **948**, 1–148 (2022)
40. D.D. Lopes, B.R. da Cunha, A.F. Martins, S. Gonçalves, E.K. Lenzi, Q.S. Hanley, M. Perc, H.V. Ribeiro, Sci. Rep. **12**, 15746 (2022)
41. E. Sayari, S.T. da Silva, K.C. Iarosz, R.L. Viana, J.D. Szezech Jr., A.M. Batista, Chaos Solitons Fractals **164**, 112616 (2022)
42. H.E. Suna, H. Tanberkan, B.S. Gür, M. Perc, M. Özer, J. Econ. Cult. Soc. **61**, 41–64 (2020)
43. K. Roster, C. Connaughton, F.A. Rodrigues, Am. J. Epidemiol. **191**, 1803–1812 (2022)
44. J. Ong, X. Liu, J. Rajarethinam, S.Y. Kok, S. Liang, C.S. Tang, A.R. Cook, L.C. Ng, G. Yap, PLoS Negl. Trop. Dis. **12**, e0006587 (2018)
45. C.M. Benedum, K.M. Shea, H.E. Jenkins, L.Y. Kim, N. Markuzon, PLoS Negl. Trop. Dis. **14**, e0008710 (2020)
46. L. Breiman, Mach. Learn. **45**, 5–32 (2001)

47. I. Sanchez-Gendriz, G.F. de Souza, I.G.M. de Andrade, A.D.D. Neto, A.M. Tavares, D.M.S. Barros, A.H.F. de Morais, L.J. Galvão-Lima, R.A.M. Valentim, Sci. Rep. **12**, 6550 (2022)
48. Instituto Nacional de Meteorologia: Ministério da Agricultura e Pecuária. https://portal.inmet.gov.br. Accessed 10 Aug 2023
49. National Oceanic and Atmospheric Administration. Dengue Forecasting Project Data Repository. https://dengueforecasting.noaa.gov. Accessed 10 Aug 2023
50. Sistema Nacional de Vigilancia en Salud Pública—Sivigila. https://portalsivigila.ins.gov.co. 10 Aug 2023
51. J.C. Trujillo, P. Howley, Environ. Behav. **53**, 69–90 (2021)
52. ClimateDengueForecast. https://github.com/ecgabrick/ClimateDengueForecast. Accessed 20 Apr 2024
53. urca: Unit Root and Cointegration Tests for Time Series Data. https://CRAN.R-project.org/package=urca. Accessed 21 Mar 2024
54. A. Cutler, D.R. Cutler, J.R. Stevens, Random forests, in *Ensemble Machine Learning*, ed. by C. Zhang, Y.Q. Ma (Springer, New York, 2012), pp. 157–175
55. A. Arsham, P. Rosenberg, M. Little, N. Engl. J. Stat. Data Sci. **1**, 46–61 (2023)
56. P. Probst, M.N. Wright, A.L. Boulesteix, Adv. Rev. **9**, e1301 (2019)
57. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. **12**, 2825–2830 (2011)
58. S. Janitza, R. Hornung, PLoS One **13**, e0201904 (2018)
59. G. Biau, E. Scornet, TEST **25**, 197–227 (2016)
60. A.J. Izenman, *Modern Multivariate Statistical Techniques* (Springer, New York, 2008)
61. D.M. Belete, M.D. Huchaiah, Int. J. Comput. Appl. **44**, 875–886 (2022)
62. Use Time Series to Generate and Compare Power Spectral Density (PDSR). https://CRAN.R-project.org/package=psdr. Accessed 1 Mar 2024