Dengue Vector Population Forecasting Using Multisource Earth Observation Products and Recurrent Neural Networks

Oladimeji Mudele, Alejandro C. Frery, *Senior Member, IEEE*, Lucas F. R. Zanandrez, Alvaro E. Eiras, and Paolo Gamba, *Fellow, IEEE*

***Abstract*—This article introduces a technique for using recurrent neural networks to forecast *Ae. aegypti* mosquito (Dengue transmission vector) counts at neighborhood-level, using Earth Observation data inputs as proxies to environmental variables. The model is validated using *in situ* data in two Brazilian cities, and compared with state-of-the-art multioutput random forest and k-nearest neighbor models. The approach exploits a clustering step performed before the model definition, which simplifies the task by aggregating mosquito count sequences with similar temporal patterns.**

***Index Terms*—Deep learning, dengue risk, remote sensing, satellite images, *Aedes aegypti*.**

I. Introduction

ZOONOTIC diseases are one of the most widespread threats to human lives in many parts of the world. Dengue is a very prevalent one of such. This disease is transmitted by the female *Ae. aegypti* mosquito species. The female *Ae. aegypti* is fully adapted to urban areas and breeds in artificial water containers. The spread of Dengue has been shown to be significantly influenced by the density of mosquito vector in any location, which itself is a result of local biotic and abiotic environmental interactions. Specifically, the environmental variables, which have shown empirical relationship with the development and population of female *Ae. aegypti* are precipitation, humidity, vegetation condition, and land surface temperature (LST) [1]–[4].

Many studies have explored ways to understand the interaction effects of these environmental variables on the diseases spread, both locally and globally. While some of these studies focus on modeling and prediction of risks [3] or vector population [5] based on environmental effects, others go further to explain these interactions so as to gain better understanding of the most important environmental driving mechanisms [2]. Regardless of the study goal, there is the need for quality and scalable data sources with which to estimate these environmental driving effects. As a result, due to their global availability and free access in many cases, Earth Observation (EO) satellite images have recently become prevalent for this application [1], [2], [6].

The baseline of this study is obtained from studies presented in [1], [2], [5], all devoted to "nowcasting" of *Ae. aegypti* mosquito population temporal distribution at the municipality level starting from environmental conditions estimated from freely available satellite image products. Here in this study, for the first time, a methodology for *Ae. aegypti* population forecast model, which is spatially disaggregated at the neighborhood level is presented. To this aim, the same freely available EO satellite image products as in [1] and [2] are used for the estimation of the environmental features of interest. Forecast models, as opposed to nowcasting models (e.g., [1], [2], [5]), serve to enable operational disease outbreak surveillance systems to anticipate and better plan for future disease spreads.

Time series forecasting methods have found applications in various domains e.g., economics [7], weather, and environmental state predictions [8]. Some of the most frequently used algorithms include autoregressive moving average model (ARMA), autoregressive integrated moving average (ARIMA) [9], random forest (RF), and, more recently, neural and deep learning networks [10]. Among other neural network approaches, recurrent neural networks (RNNs) [11], [12] have been used for sequential data modeling, and show great capability to capture multivariate nonlinear interactions in data sequences [13]. Due to RNNs' capability to capture long-term temporal dependencies in input data, they have proven to attain better quality than traditional feed-forward neural networks for the specific purpose of time series modeling [14]. However, vanilla RNNs suffer the problem of vanishing and exploding gradients over long sequences [9]. As a result, long short-term memory (LSTM) [15] and gated recurrent unit (GRU) [16], which are variants of RNN designed

Manuscript received November 4, 2020; revised February 6, 2021 and March 20, 2021; accepted April 3, 2021. Date of publication April 14, 2021; date of current version May 7, 2021. This work was supported in part by the European Commission through Horizon 2020 Research and Innovation Programme, under Grant 734541 - Project EOXPOSURE. *(Corresponding author: Oladimeji Mudele.)*

Oladimeji Mudele and Paolo Gamba are with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia PV, Italy (e-mail: oladimeji.mudele01@universitadipavia.it; paolo.gamba@unipv.it).

Alejandro C. Frery is with the School of Mathematics and Statistics, Victoria University at Wellington, Kelburn, Wellington 6012, New Zealand (e-mail: alejandro.frery@vuw.ac.nz).

Lucas F. R. Zanandrez is with the Ecovec Ltda, Belo Horizonte 31310-260, Brazil (e-mail: lucasfabrini93@gmail.com).

Alvaro E. Eiras is with the Laboratory of Technological Innovation and Entrepreneurship in Vector Control Department of Parasitology, Federal University of Minas Gerais, Belo Horizonte 31270-901, Brazil (e-mail: alvaro.eiras@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3073351

to mitigate the earlier stated setbacks, were designed. LSTM and GRU have found successful application in many fields, *e.g.*, machine translation [17], speech recognition [18], and other time series forecasting tasks [10]. Recent state-of-the-art time series forecasting applications consider the use of LSTM and GRU in an encoder–decoder fashion [10]. Consequently, RNNs (LSTM and GRU) could also be applied to the problem of epidemiological forecasting.

Accordingly, the research question in this work is whether an accurate geographically distributed time series prediction for *Ae. aegypti* numbers at the neighborhood level is possible using EO data as inputs to RNNs. This study looks to tackle the question of obtaining a qualitative time series prediction of the population of female *Ae. aegypti* at neighborhood-level based on EO data inputs to RNNs. To address this question, we refer to the study presented in [19], which shows that in a group of concurrent mosquito population time series data covering a specific area, and over a sufficient amount of time, there exist multiple subgroups (or clusters) of temporally homogeneous time series in different spatial points. From the results of that study, we deduce that the temporal distribution of mosquito population data within the same cluster can be approximated as a single signal: the centroid (or mean) of this cluster. Leveraging this technique, the problem of neighborhood-level *Ae. aegypti* vector population modeling has been split into two steps: (i) finding vector population time series cluster along the spatial axis (similar time series signals) and obtaining their mean signals (centroids); and (ii) deriving a model of the obtained means using environmental information at the neighborhood-level from free EO products. Point (i) is achieved using *k*-means clustering, and point (ii) using RNNs. The innovation of this study is that it introduces a neighborhood-level one-week-ahead vector population forecasting technique, while other studies in this domain only either perform regional-level [3] or urban-level [1], [2], [5] nowcasting. Additionally, the use of RNNs along with EO data features has not been considered in this domain, as far as we know.

The structure of this article is as follows: the next section is devoted to providing some background on the use of EO data for dengue risk modeling, while Section III provides some background on RNNs. The methodology proposed in this work is presented in Section IV, and the datasets (both spaceborne and *in situ*) used in this work are introduced. The experimental results are shown in Section VI, while Section VII is devoted to discussing the models abilities and limitations. Finally, Section VIII concludes this article.

II. Background on the Use of Satellite Image Products for *Ae. Aegypti* Monitoring

There is significant evidence of the effects of temperature, humidity, precipitation and surface vegetation on the life-cycle, development and density of female *Ae. aegypti* mosquito species in urban environments [3]. The air temperature affects the species survival as well as the duration of the extrinsic incubation period of the virus and the length of the gonotrophic cycle [20].

Since the *Ae. aegypti* larvae and pupae breed mostly in artificial containers in urban areas, precipitation serves as a source of water for such containers. Many studies [21]–[23] have shown that there is positive correlation between Dengue risk infection and precipitation levels. Furthermore, since the *Aedes* mosquitoes lay eggs in areas of high moisture [24], higher relative humidity has been shown to affect propagation of female adult *Ae. aegypti* vectors locally [6].

Vegetation condition is another important variable in spatio-temporal behavior of the vector population cycle. For example, vegetation canopy cover has been associated with larger larvae density because it reduces evaporation from breeding containers, it decreases subcanopy wind speed, and it protects outdoor habitats from direct sunlight [3]. Also, the form of organic content (plants and insects they host) that fall in breeding containers affect the development of mosquito larvae [25].

To account for the all these environmental effects, the studies in [1], [2], and[5] used the normalized difference water index (NDWI) [26]—as proxy for humidity. NDWI compares the reflectance in near-infrared and mid-infrared channels of optical satellite images. These same studies used the normalized difference vegetation index (NDVI) [27], which is a measure of surface chlorophyll level, to represent vegetation effect. NDVI compares the satellite data reflectance values in red and near-infrared channels. Other studies [3] have used the enhanced vegetation index [28], which gives more information about the vegetation canopy structural variation than NDVI.

Additionally, some optical EO satellite data products come with thermal infrared bands. These bands can be used to obtain the LST through a range of techniques, which take atmospheric effects and surface emissivity into account. LST layers have been used in [1], [2], and [5] to account for temperature effects in vector development.

In technical literature, most studies exploit NASA's moderate resolution imaging spectroradiometer (MODIS) data products to obtain indices and layers representing NDVI, NDWI, and LST information. Precipitation, on the other hand, is accounted for by using data obtained from the integrated multisatellite retrievals for global precipitation measurement (IMERG) technique [29]. IMERG is an algorithm used to intercalibrate, merge, and interpolate spaceborne microwave precipitation estimates, together with microwave-calibrated infrared satellite estimates, precipitation gauge data, and potentially other precipitation estimators at fine time and space scales over the entire globe. IMERG derived satellite data products include the tropical rainfall measuring mission data, and global precipitation measurement data [30].

III. Background on RNNs

Unlike feed forward neural networks, RNNs are a kind of neural networks with loops, which allow them to learn sequential dependency in data. Given $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ with $\mathbf{x}_t \in \mathbb{R}^u$ as input independent covariate features, a simple RNN can be expressed as follows:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t) \tag{1}$$

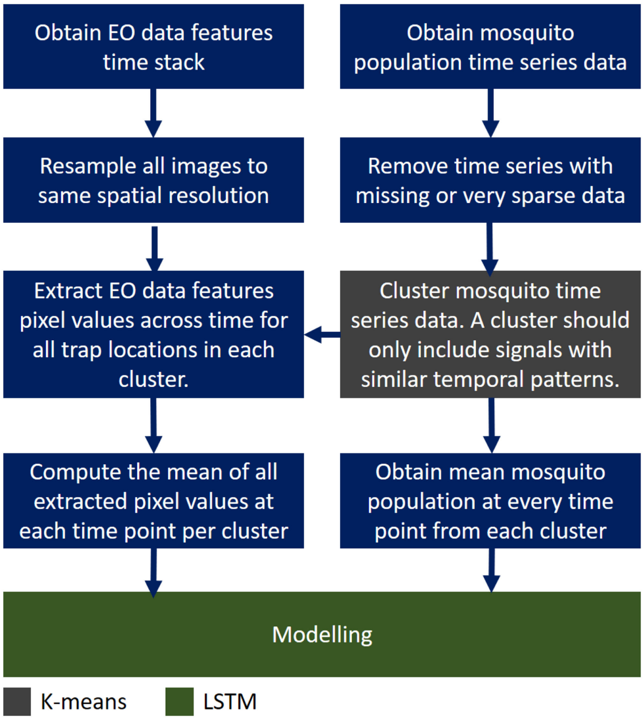

Obtain EO data features time stack
Obtain mosquito population time series data
Resample all images to same spatial resolution
Remove time series with missing or very sparse data
Extract EO data features pixel values across time for all trap locations in each cluster.
Cluster mosquito time series data. A cluster should only include signals with similar temporal patterns.
Compute the mean of all extracted pixel values at each time point per cluster
Obtain mean mosquito population at every time point from each cluster
Modelling
K-means
LSTM

Fig. 1. Schematic of the study methodology.

where $\mathbf{h}_t \in \mathbb{R}^v$ is the hidden state at time t, and v is the number of hidden units, which is an hyperparameter to set.

Due to the problem of vanishing gradients RNNs, the function f is estimated using LSTMs [15]. An LSTM maintains a hidden state, $\mathbf{h}_t$, and a memory cell state, $\mathbf{s}_t$, that are updated at every time step, and used to determine the output at that same time. At each time step, the access to s_t is controlled by three sigmoid gates: forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, and output gate $\mathbf{o}_t$. The mathematical formulations of these gates and the resulting $\mathbf{h}_t$ and $\mathbf{s}_t$ are summarized as follows:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1};\mathbf{x}_t] + \mathbf{b}_f)\\
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{h}_{t-1};\mathbf{x}_t] + \mathbf{b}_i)\\
\mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{h}_{t-1};\mathbf{x}_t] + \mathbf{b}_o)\\
\mathbf{s}_t &= \mathbf{f}_t \odot \mathbf{s}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_s[\mathbf{h}_{t-1};\mathbf{x}_t] + \mathbf{b}_s)\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{s}_t)
\end{aligned} \tag{2}$$

where $[\mathbf{h}_{t-1};\mathbf{x}_t] \in \mathbb{R}^{v+u}$ is a concatenation of the previous hidden state $\mathbf{h}_{t-1}$ and the current input $\mathbf{x}_t$; $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$, $\mathbf{W}_s$ $\in \mathbb{R}^{v\times(v+u)}$, and $\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_o, \mathbf{b}_s \in \mathbb{R}^m$ are learnable weight and bias parameters, respectively; σ, $\tanh$, and $\odot$ are the logistic sigmoid activation function, the hyperbolic tangent function, and the Hadamard product, respectively.

IV. Methodology

Fig. 1 shows a high-level schematic of the study methodology. As shown by this figure, the mosquito time series data are first partitioned into clusters using a k-means clustering technique. The resulting cluster centers are then modeled using RNN with EO data as input features.

In the following sections, details about the RNN architecture and K-means clustering are provided.

A. Notation and Problem Statement

Let's consider a database of concurrent time series of *Ae. aegypti* mosquito population collected using M mosquito traps: $Y = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(M)}\}$, with $\mathbf{y}^{(m)} \in \mathbb{R}^P$, where P is the observation period (e.g., the total number of weeks in case of weekly monitoring), $\mathbf{y}^{(m)}$ is the vector of data collected at the mth mosquito trap, i.e., $\mathbf{y}^{(m)} = (y_1^{(m)}, y_2^{(m)}, \ldots, y_P^{(m)})$. Instead, let us denote $\mathbf{y}_t = (y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(M)})$ the vector of mosquito numbers collected at all M traps in a particular tth week.

Now, let us partition Y into K clusters of time series $C = \{C^{(1)}, C^{(2)}, \ldots, C^{(K)}\}$ with means (i.e., cluster centroids) $\{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(K)}\}$, $\mathbf{c}^{(k)} \in \mathbb{R}^P$. The clusters form a partition of Y, i.e., $C^{(k)} \subseteq Y, (k = 1, 2, 3, \ldots, K), \cap_{k=1}^{K} C^{(k)} = \emptyset$ and $\cup_{k=1}^{K} C^{(k)} = Y$. We present details of the clustering algorithm in Section IV-C.

Finally, let us consider N environmental variables (or proxies to them extracted from EO data), whose measures are available for the same P time instants in the M locations of the mosquito traps. Let us denote the whole set of values of these variables as $\mathbf{V} \in \mathbb{R}^{N\times M\times P}$. By clustering $\mathbf{V}$ according to the partition of Y, $\mathbf{V}$ is reduced to $\mathbf{X} \in \mathbb{R}^{N\cdot K\times P}$, where $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N\cdot K)}\}$, and $\mathbf{x}^{(i)} \in \mathbb{R}^P$. Eventually, let $\mathbf{x}_t \in \mathbb{R}^{N\cdot K}$ be the set of mean values of the environmental variables for each cluster at the time instant t, i.e., $\mathbf{x}_t = \{x^{(1)}(t), x^{(2)}(t), \ldots, x^{(N\cdot K)}(t)\}$.

Using a temporal window of size $T \ll P$, we formulate our forecast model as a nonlinear autoregressive exogenous model (NARX) as follows:

$$\widehat{\mathbf{c}}_t = F([\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}]; [\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}]) \tag{3}$$

where ";" denotes the time point concatenation, and, as before, $k = 1, 2, \ldots, K$. $F(.)$ is selected to be an LSTM model (see Section IV-B. The model output $\widehat{\mathbf{c}}_t \in \mathbb{R}^k$ is a vector of the forecast values of mean mosquito population for the k clusters in the tth week based on T trailing autoregressive (vector population) and exogenous (environmental conditions) components.

B. Adaptation of RNNs for This Work

For this study, we used an encoder–decoder LSTM [31] architecture due to its recorded success in many applications, including time series forecasting. The encoder is an LSTM, which encodes the input sequence, $\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}$ and $\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}$ within a time window of length T, into a learned representation $\mathbf{h}_t \in \mathbb{R}^v$ and memory cell state $\mathbf{s}_t \in \mathbb{R}^v$, where v is the encoder output size. For time series prediction tasks, the decoder is usually a stack of LSTM and a fully connected (dense) neural network layer with nonlinear activation. The decoder LSTM takes $\mathbf{h}_t$ as input, copies it over the length of T, and generates the decoder hidden state $\mathbf{d}_t$. The fully connected layer takes $\mathbf{d}_t$ as input and produces $\widehat{\mathbf{c}}_t$. For this study, we added a fully connected layer with a rectified linear activation function (ReLU) [32] on

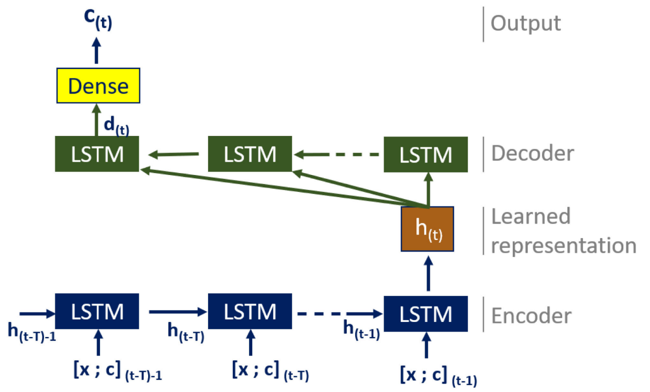


Fig. 2. Architecture of our adapted encoder–decoder LSTM. The encoder output $\mathbf{h}_t$ is replicated into T copies to feed each time point of the decoder. The dense layer maps the decoder output to the desired prediction output. ";" signifies concatenation; T is the size of a temporal window; $\mathbf{c}_t$ is the predicted output vector at time t; $\mathbf{x}_{t-1}$ is a vector of the EO covariate features at time $t-1$.

top of the decoder LSTM to map the output of the LSTM to a vector of predicted mosquito populations. Fig. 2 illustrates this adapted encoder–decoder LSTM.

Considering a window of size T and subwindows of size P such that $T \ll P$ and $T \mod P = 0$, the model is

$$\mathbf{h}_t = f_1([\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}]; [\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}]) \quad (4)$$

$$\mathbf{d}_t = f_2(\mathbf{h}_t) \quad (5)$$

$$\widehat{\mathbf{c}}_t = \vartheta(\mathbf{W}_d \mathbf{d}_t + \mathbf{b}_d) \quad (6)$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are learnable parameters of the decoder fully connected layer, and $\mathbf{d}_t$ and $\widehat{\mathbf{c}}_t$ are the decoder hidden state and model output for time t prediction, respectively; $f_1(.)$ and $f_2(.)$ are the encoder and decoder LSTMs, respectively; ϑ is the ReLU activation function, which is defined for an arbitrary input $x \in \mathbb{R}$ as $\vartheta(x) = \max\{0, x\}$.

The choice of the ReLU activation on our modeling output is justified by the need to produce positive real number output predictions i.e., $\widehat{c}_t^{(i)} \in \mathbb{R}^+ \; \forall \; \widehat{c}_t^{(i)} \subseteq \widehat{\mathbf{c}}_t$, since it is impossible to have negative mosquito vector population values.

C. Time Series Clustering

The clustering applied to the set of mosquito trap records Y is implemented by means of the standard *K*-means algorithm with Euclidean distance. This algorithm is simple to implement and converges fast [33], [34]. Still, due to the unsupervised nature of clustering, there is the need to determine the optimal number of clusters K. The goal is selecting K to minimize the total intracluster variation, also known as total within-cluster sum of square variation or distortion. To this aim, we used the elbow method [35]. The resulting set of distortions is then plotted, the "optimal" K is selected as the *"sweet spot,"* where there is a bend ("elbow") in the curve indicating a significant reduction in the gradient of the distortion with respect to K. The distortion is computed as follows:

$$J = \sum_{k=1}^{K} \sum_{m=1}^{M} \left\| \mathbf{y}^{(m)} - \mathbf{c}^{(k)} \right\|^2. \quad (7)$$

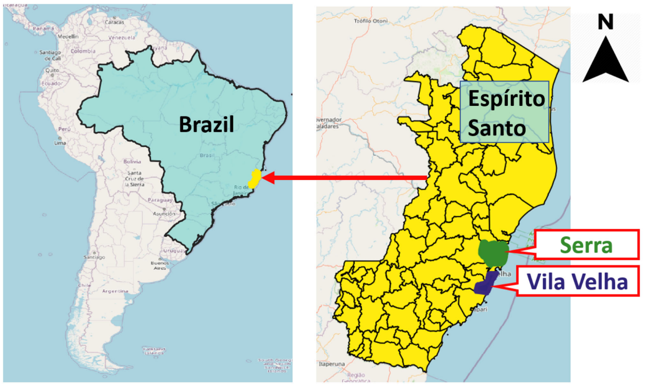


Fig. 3. Geographical location of the considered study areas: Vila Velha and Serra.

V. Materials

A. Study Area and Field Data

This research is based on adult *Ae. aegypti* mosquito counts collected in the cities of Vila Velha and Serra in Espírito Santo State (region), Brazil. Vila Velha is between latitudes $20°19'$ and $20°32'$ South, and longitudes $40°16'$ and $40°28'$ West. It covers a total area of $209.965\ \text{km}^2$, and has an estimated population of 486 208 people. Serra is between latitudes $20°7'$ and $20°12'$ South, and longitudes $40°18'$ and $40°30'$ West. It covers a total area of $553\ \text{km}^2$, and has an estimated population of 507 598 people. Both cities are about $40\ \text{km}$ apart. Fig. 3 presents their geographical locations.

These chosen locations are very relevant to dengue risk mapping. Specifically, in 2019, the Espírito Santo state verified 63 847 dengue cases with an incidence of 1588.8 per 100 000 inhabitants [36]. Vila Velha had 6611 of those cases (1359.7.6 per 100 000 inhabitants), which is far greater than the epidemic threshold.

Mosquito counts were collected weekly from MosquiTRAP devices, which are sticky mosquito traps designed to directly measure adult *Ae. aegypti* abundance and generate data for targeting vector control activities. Mosquito abundance and counts are sent in near real-time to an online surveillance system, called MI-Aedes [37], which was implemented in Vila Velha and Serra from 2017.

In Vila Velha, the record spans from Apr. 10, 2017 to Oct. 5, 2018 coming from 791 MosquiTRAPs, whereas in Serra, the record spans from Apr. 27, 2017 to Dec. 06, 2018 and come from 1127 devices. All devices in both locations are placed at least at 250 m apart. Data were acquired on site weekly by a team of trained and supervised field workers by inspecting the sticky cards set inside each trap. *Ae. aegypti* specimens were identified, counted, and their presence and number registered thanks to a mobile app.

TABLE I
DETAILS OF EO DATA PRODUCTS USED AND THE PROXY ENVIRONMENTAL VARIABLE THEY REPRESENT

Data product	Band(s) used	Feature	Spatial resolution	Temporal resolution	Proxy to:
			(meters)	(days)	
MODIS MOD11A2	`LST_Day_1km`	Daytime LST [a]	1000	8	Maximum temperature
MODIS MOD11A2	`LST_Night_1km`	Night-time LST	1000	8	Minimum temperature
GPM	`precipitationCal`	Precipitation	≈ 11,000	1	Precipitation
MODIS MOD13Q1	`NDVI`	NDVI [b]	250	16	Vegetation condition
MODIS MOD13Q1	`sur_refl_b02` and `sur_refl_b07`	NDWI [c]	250	16	Surface moisture and humidity

[a] LST: Land surface temperature [38]. [b] NDVI: Normalized Difference Vegetation Index (NDVI) [27]. [c] NDWI: Normalized Difference Water Index (NDWI) [26].

As a result of the control activities optimization performed at the start of year 2018, the data collected are divided into two temporal regimes: 2017 and 2018. In Vila Velha, the data for 2017 spans from Apr. 10, 2017 to Dec. 31, 2017 (epidemiological weeks 15–52, 36 weeks), while the data for 2018 spans from Jan. 02, 2018 to Oct. 05, 2018 (epidemiological weeks 1–40, 40 weeks). In Serra, the data for 2017 spans from Apr. 27, 2017 to Dec. 30, 2017 (epidemiological weeks 17–52, 36 weeks), while the data for 2018 spans Jan. 05, 2018 to Oct. 05, 2018 (epidemiological weeks 1–49, 49 weeks). Note that weeks 7 and 8 for year 2018 are not available because Serra's field workers temporarily deactivated traps in week 7 due to the Carnival holiday in week 8. For the purposes of completeness, epidemiological weeks are simply a standardized method of counting weeks to allow for the comparison of data year after year, especially in cases, where the data are not seasonally aligned.

To preprocess the data, traps missing data even for just one of the weeks or with zero mosquito reported in all weeks were filtered out. This was done to avoid the creation of synthetic *in situ* data through interpolation. This filtering resulted in a final set of 193 and 325 trap records in 2017 and 2018, respectively, out of the initial 791 points in Vila Velha. Similarly, in Serra, the final set includes 567 trap records in 2017 and 95 in 2018. Just to reduce random "data noise" in the obtained series from each retained trap, an exponential moving average filter with a span of five weeks was applied to the records.

B. Environmental Variables From EO Data

Table I describes the EO products used in this study together with the specific bands and resolutions (spatial and temporal), as well as the environmental variable to which they act as proxy.

All the utilized satellite data products were accessed using the `Javascript` application programming interface (API) of Google Earth Engine. Specifically, the data were downloaded using the `export` method of this API. This method has resampling and reprojection functions wrapped into it for easy coregistration of multisource data with different properties. All datasets were obtained at 250 m spatial resolution by nearest-neighbor resampling because this is the minimum distance separating neighboring mosquito traps [2]. All the necessary extraction and processing steps applied to these data are presented in details in the next section.

C. Data Extraction and Transformation

Since the approach is supervised, there is the need to select training and validation samples. In this research, this step was performed after the clustering, because the forecast model is applied [(3)] to the cluster representative values.

Therefore, for each cluster, the average values of the environmental covariate features at each point in time ($\mathbf{x}_t$) were computed by averaging the EO proxy values in the locations of the traps assigned to that cluster. Variables representing different time intervals were temporally interpolated with a third-order spline to obtain weekly values.

In parallel, the mosquito count records for each cluster were randomly subdivided into two sets: one for training, and the other one for model testing. Accordingly, the $\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}$ vectors in (3), were estimated using only the training set in the training phase, and the test set in the model testing phase.

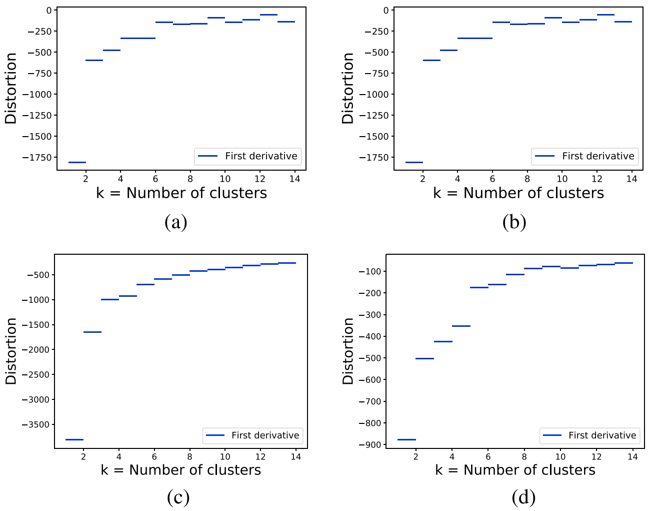
Finally, the resulting training data (target and predictor variables combined) is then randomly subdivided along time to extract 20% of the time-points to be used for validation of the model during training.

VI. EXPERIMENTAL RESULTS

A. Training Procedure

The model was trained for one-week-ahead female *Ae. aegypti* population prediction starting from T training populations (autoregressive component) and environmental condition features (exogenous component). As a result, we obtained predictions starting at $t = T + 1$.

The adaptive learning rate optimization algorithm (Adam) [39] was selected to train the neural network with a learning rate of 0.001. The objective function for parameters learning through backpropagation was set to the mean absolute error (MAE) loss [10]. A dropout rate of 0.2 was used in the decoder to avoid overfitting, and a batch size of 1 because the dataset is not too large. All the models were trained in 100 epochs, and the model with the best validation accuracy was selected and saved.



(a) (b) (c) (d)

Fig. 4. Numerical first derivatives of the elbow plots for selecting the optimal number of *k*-means clusters in 2017 and 2018. The plots show that $k = 6$ is the elbow point in both years in Vila Velha, while $k = 5$ is the elbow point in both years in Serra. (a) vila velha, 2017. (b) vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

B. Parameter Settings

The three key parameters required in an encoder–decoder LSTM are: 1) the temporal window size T, 2) the encoder $\mathbf{h}$ output vector size, and 3) the decoder $\mathbf{d}$ output vector size. For simplicity, as in [10], both the encoder and the decoder ($\mathbf{h} \in \mathbb{R}^v \ni \mathbf{d}$) use a single layer each, i.e., their output size is v. We considered the window size T as one of the three possible values {3, 6, 9 }, while the value of v is selected from the set {16, 32, 64, 128 }.

C. Baseline Models for Comparison

To prove their significance, the results obtained by LSTM will be compared to those obtained from other regression models. The models are as follows:

1) RF: an ensemble learning method consisting of a collection of several randomized regression trees. For this study, a multioutput RF model is fitted with the number of trees set to the commonly selected value of 500 (e.g., as in [2]).
2) k-nearest neighbor regression (KNN): a multioutput regression based on the 5-nearest neighbors using the Euclidean distance metric. The method has been chosen because it produced the best result in [1].

These models are fitted according to the NARX introduced in (3). However, due to the nonrecurrent nature of both RF and KNN, it is not possible to consider the sequential ordering of lagged environmental effects within the considered time window T. Hence, they are concatenated into a single vector, ignoring their temporal ordering.

D. Clustering Results

To address the stability problem of the *K*-means clustering method, the elbow plot is obtained as the average of 20 repetitions over the sequence $1 \leq K \leq 14$. Fig. 4 shows the piecewise approximated first derivatives of the resulting *K*-means elbow plots in Vila Velha [Fig. 4(a) and (b)] and Serra [Fig. 4(c) and (d)] in both years. In Vila Velha, beyond $K = 6$, there is no significant jump in the derivative, hence we chose $K = 6$ as optimal elbow point in that location in both years. For Serra, on the other hand, $K = 5$ is the optimal elbow point, which is chosen for both years. The clusters are derived and labeled such that the six clusters in Vila Velha are labeled Cluster 1A–6A in 2017 and Cluster 1B–6B in 2018. The A and B suffixes of the cluster labels are codes for years 2017 and 2018, respectively. Following the same convention, Cluster 1A–5 A and 1B–5B representing the five clusters for both years in Serra are also derived.

The line plots of Fig. 5 show the mosquito population temporal patterns for each resulting cluster for each year and location. Different clusters are differentiated by their temporal distribution and range (see the *y*-axes of plots). In an epidemiological sense, periods of maximum spikes are indicative of highest possible disease outbreak risks. In each case presented in Fig. 5, the *K*-means clustering has helped to identify underlying common patterns that describe the vector development activity at the different sites, where trap observations have been carried out.

Furthermore, it is seen from Fig. 5 that in spite of the intercluster temporal pattern differences, there are similar patterns in subsequences across multiple clusters and locations. Such similarities correspond to weeks of typical macroclimatic effects at municipal and regional levels. By the hypothesis of vector population dependency on abiotic and biotic environmental effects, differences in temporal patterns of the cluster centers correspond to differences in microclimatic effects, which differ across clusters, and are shared within the same cluster. A microclimate is defined by a set of atmospheric conditions that differ from those in the surrounding areas.

In Vila Velha, in 2017, local peaks can be observed in the neighborhood of epidemiological weeks 25, 35, 43, and 52, respectively. For each cluster, however, the range and duration of the peaks differ. In 2018, the population of the mosquitoes is always decreasing between observation (and epidemiological weeks) 1–9 for all clusters. There are spikes with local peaks in the neighborhood of week 25 in clusters 3 and 6. For Serra, in 2017, we have local peaks in the neighborhood of epidemiological weeks 34, 44, and 48, respectively, for all clusters.

Common patterns are also observed among some subsets of clusters. For example, all clusters except 2A exhibit increasing vector population between observation weeks 6–16. Still in Serra, in 2018, Clusters 4B and 5B exhibit different patterns all through the year with respect to the other clusters that are always close to zero. Intercluster similarities per location can be attributed to municipality-level macroclimatic effects.

We also see patterns that are common to both Vila Velha and Serra. For example, in 2017 there are local peaks in the neighborhood of epidemiological weeks 34–36 and 43–44 in both test locations. These similarities can be attributed to regional macroclimatic effects. Similarities—at municipality and regional levels—only exist in pockets of time duration, as shown in Fig. 5. This result reveals the strengths and weaknesses of municipality-level modeling like the one obtained in [1], [2],

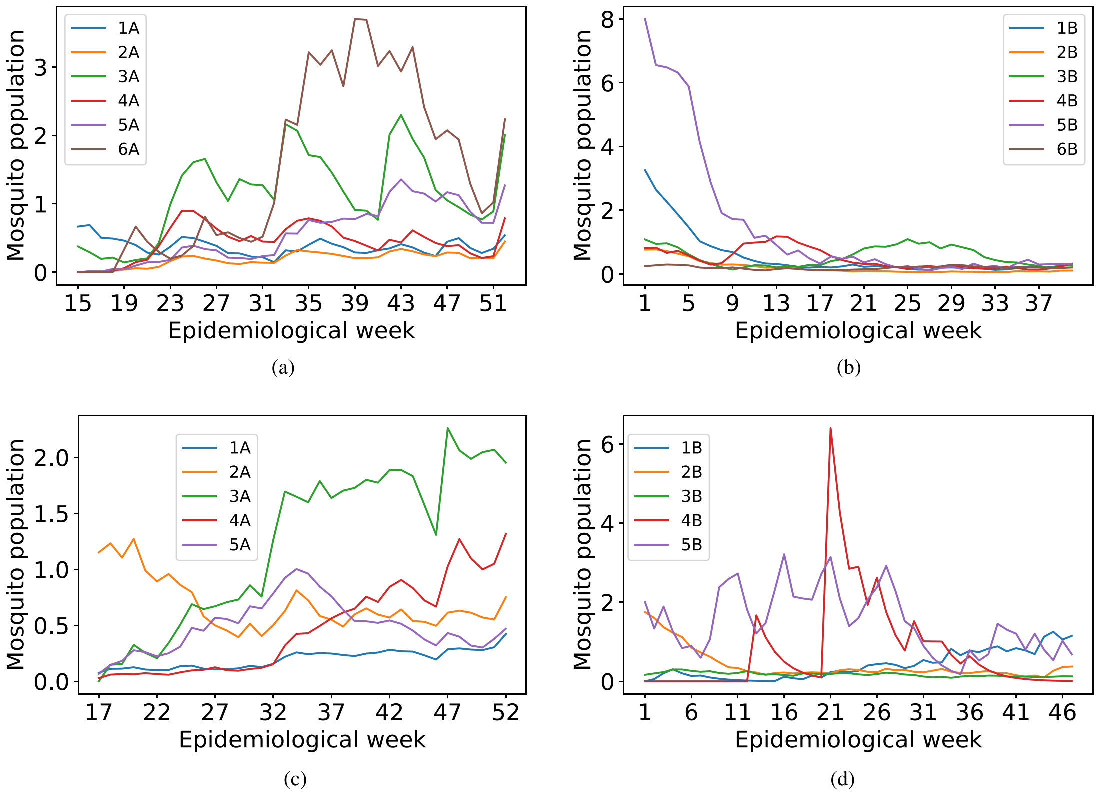


Fig. 5. Mean temporal distribution for clusters obtained in both years. The *y*-axis represents the cluster mean mosquito population (**c**) while the *x*-axis represents the epidemiological week. (a) Vila Velha, 2017. (b) Vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

and [5], especially for [2], which uses the same data in Vila Velha as in this study. While such municipality-level models can capture general trends that are common to most clusters, they do not provide detailed information across different clusters. In areas, where the trends in all clusters are similar, then, perhaps, a municipality-level model can be sufficient. Otherwise, there is the need for a disaggregated approach like the one presented in this study for better inference at neighborhood-level.

Fig. 6 presents boxplots of the vector population series for the derived clusters in both test locations. In the epidemiological sense, minima, maxima, and interquartile ranges (IQR) provide a risk profile summary of the component points of each the cluster. Higher maxima mean higher risk exposure at peak periods, while the minima are the lower bounds of the risk exposure in the locations considered. The IQRs describe the pattern variability of the risk exposure in the considered time. In Vila Velha, Cluster 6A has the highest maximum and variability in 2017, which Clusters 1A and 2A come from low risk locations. In 2018, cluster 5B has the highest maximum and variability. In Serra, Cluster 3A has the highest maximum in 2017, while cluster 5B has the largest IQR maximum in 2018. Also, Cluster 2A has the highest minimum in this same location in 2017. These intercluster differences in the vector population suggest that the clustering process has achieved the useful aim of finding homogeneous trap locations: separating the trap points into clusters of different temporal patterns and disease risk profiles.

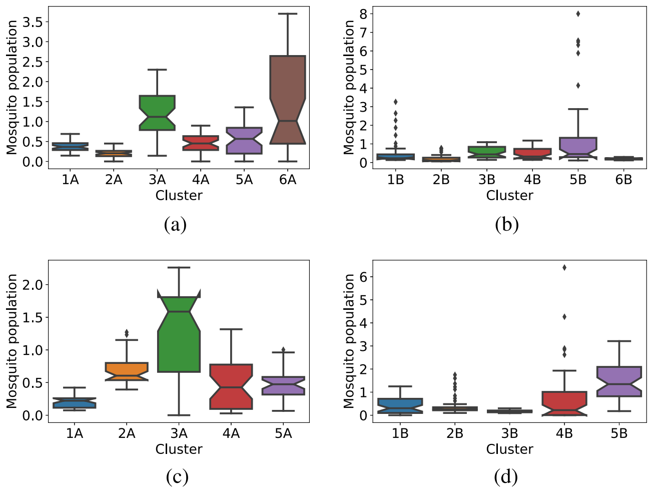
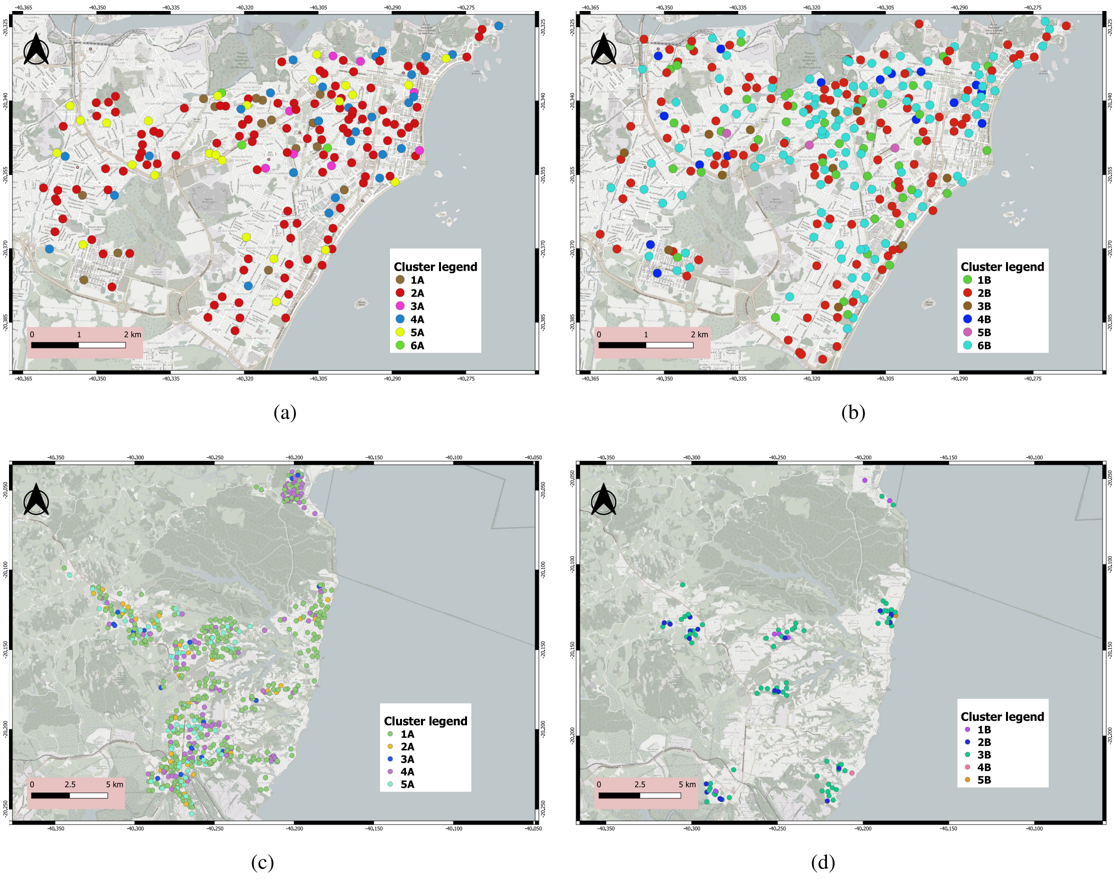


Fig. 6. Boxplots of the resulting female *Ae. aegypti* traps data cluster means. (a) Vila velha, 2017. (b) Vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

Fig. 7 presents the location of the trap points along with color indicators showing the clusters they belong to. It can be seen that points in the same cluster are not necessarily geographically collocated, as is also the case in the results presented in [19].

We examined cluster membership of points common to both years to understand the spatial relationship between clusters obtained in different years in the same location. For this, we used the overlap coefficient (OC) [40] to measure spatio-temporal similarities, i.e., the number of common traps contained between every possible cluster pair across both years in the same location.



(a)

(b)

(c)

(d)

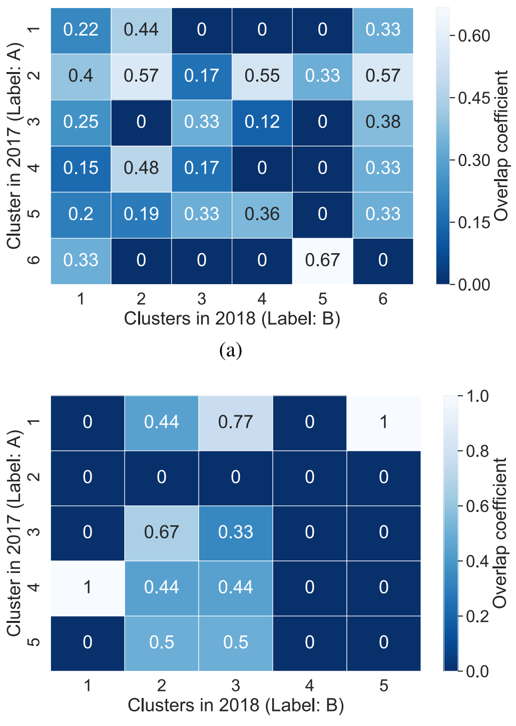
Fig. 7. Mosquito trap points color-labeled according to their clusters. In the background is OpenStreetMap$^{\text{TM}}$ view of the study areas: Vila velha and Serra. (a) Vila velha, 2017. (b) Vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

The OC is defined as follows:

$$\mathrm{OC}(C_A^i, C_B^j) = \frac{|C_A^i \cap C_B^j|}{\min(|C_A^i|, |C_B^j|)} \tag{8}$$

where C_A^i and C_B^j are the sets of mosquito trap points in the ith and jth clusters in 2017 and 2018, respectively, in the same location, after filtering all clusters to retain only traps that exist in both years. In Vila Velha, among the 193 and 325 trap points analyzed in 2017 and 2018, respectively, there are 128 traps which are common to both years. In Serra, among the 567 and 95 trap points analyzed in 2017 and 2018, respectively, there are 59 traps, which are common to both years.

The spatio-temporal similarities are presented in form of similarity matrices in Fig. 8. The results for Vila Velha [c.f. Fig. 8(a)] are discussed in the rest of this paragraph. The highest risk clusters in both years in this location—Clusters 6A and 5B—have an OC of 0.64, which is the highest similarity value obtained in the matrix. This is evidence that there is high correlation between the set of traps with high risk in both years. Also, as shown by their zero OC values with four out of the remaining five clusters, these annual highest risk clusters are decoupled from the lower risk clusters. From an epidemiological standpoint, this is evidence of continuity in risk levels across different control regimes, also showing that the microclimatic effects that drive the local vector population at these high risk



(a)

(b)

Fig. 8. Matrix of similarity in set of trap points contained in cluster pairs (each from different years). The similarity is measured by OC. (a) Vila velha. (b) Serra.

TABLE II
COMPARISON BETWEEN MAE LOSS FOR ALL MODELS WITH RESPECT TO THE TEMPORAL WINDOW SIZE WITH A CONSTANT LEARNED REPRESENTATION VECTOR SIZE; $v = 16$ IS THE LEARNED REPRESENTATION SIZE, WHILE T IS THE TEMPORAL WINDOW SIZE CONSIDERED FOR EACH PREDICTION

		Vila Velha			**Serra**		
	$T \Rightarrow$	3	6	9	3	6	9
2017	Training	0.3117	0.3392	0.4926	0.2254	0.1509	0.2451
	Validation	0.4627	0.1810	0.3745	0.1985	0.3275	0.2729
	Test	**0.6120**	0.6450	0.7565	**0.4048**	0.4703	0.5889
2018	Training	0.1407	0.2067	0.2762	0.2738	0.7999	0.2642
	Validation	0.1998	0.4516	0.3395	0.2407	0.8574	0.3151
	Test	**0.3600**	0.4624	0.4602	**0.4418**	0.9028	0.5372

TABLE III
COMPARISON OF MAE FOR MODELS WITH RESPECT TO VARYING LEARNED REPRESENTATION VECTOR SIZE FOR $T = 3$

		Vila Velha		**Serra**	
v	Year $\Rightarrow$	2017	2018	2017	2018
16	Training	0.3117	0.1407	0.2254	0.2738
	Validation	0.4627	0.1998	0.1985	0.2407
	Test	0.6120	**0.3600**	0.4048	0.4418
32	Training	0.2637	0.1501	0.1880	0.2652
	Validation	0.4774	0.1975	0.2456	0.1817
	Test	0.6274	0.4126	0.4231	**0.3986**
64	Training	0.2841	0.1781	0.1630	0.2459
	Validation	0.4765	0.2231	0.1472	0.3632
	Test	0.6203	0.3816	**0.3984**	0.4329
128	Training	0.2802	0.1808	0.2266	0.6732
	Validation	0.3880	0.3189	0.2244	0.8454
	Test	**0.5767**	0.4038	0.4440	0.5794

The learned representation is the encoder output; v: Learned representation size.

points exhibit some robustness to the control measures that have been applied.

In addition, still in Vila Velha, Cluster 2A—the one with lowest risk—has its highest OC of 0.57 with both Clusters 2B and 6B, and its lowest OC of 0.17 with Cluster 3B. As shown in Fig. 6(b), these clusters (Clusters 2B and 6B) are always low risk throughout the observation period, indicating that that they contain significant amount of the low risk points from Cluster 2A. Also, since Cluster 3B in Vila Velha is the second highest risk cluster considering the IQR, its low intersection with Cluster 2A (Vila Velha's lowest risk cluster in 2017) is in line. Clusters 1A and 2B, both of relatively low risk in both years, have an OC of 0.44.

In Serra, there are only 59 common traps points in both years. The similarity matrix is subsequently sparse [see Fig. 8(b)]. A significant amount of the sparse relationships in the matrix involve Clusters 5B and 6B, which contain only one trap each and are, thus, unreliable for the kind of analysis conducted here. In spite of the sparseness of the matrix, we still see that Clusters 1A and 3B—the lowest risk clusters in both years—have an OC of 0.77. Also, Clusters 3A and 2B have an OC of 0.67. Since Cluster 2B is a high risk cluster in 2018, if we ignore 4B and 5B, which contain single trap points each, there is again a coupling between high risk points across years and control regimes. These results further point to evidences of continuity in the risk level of the trap points even across different control regimes. Key actors in vector surveillance and control can use the information provided by this similarity matrix for neighborhood-level understanding of control activities effects.

E. Model Results

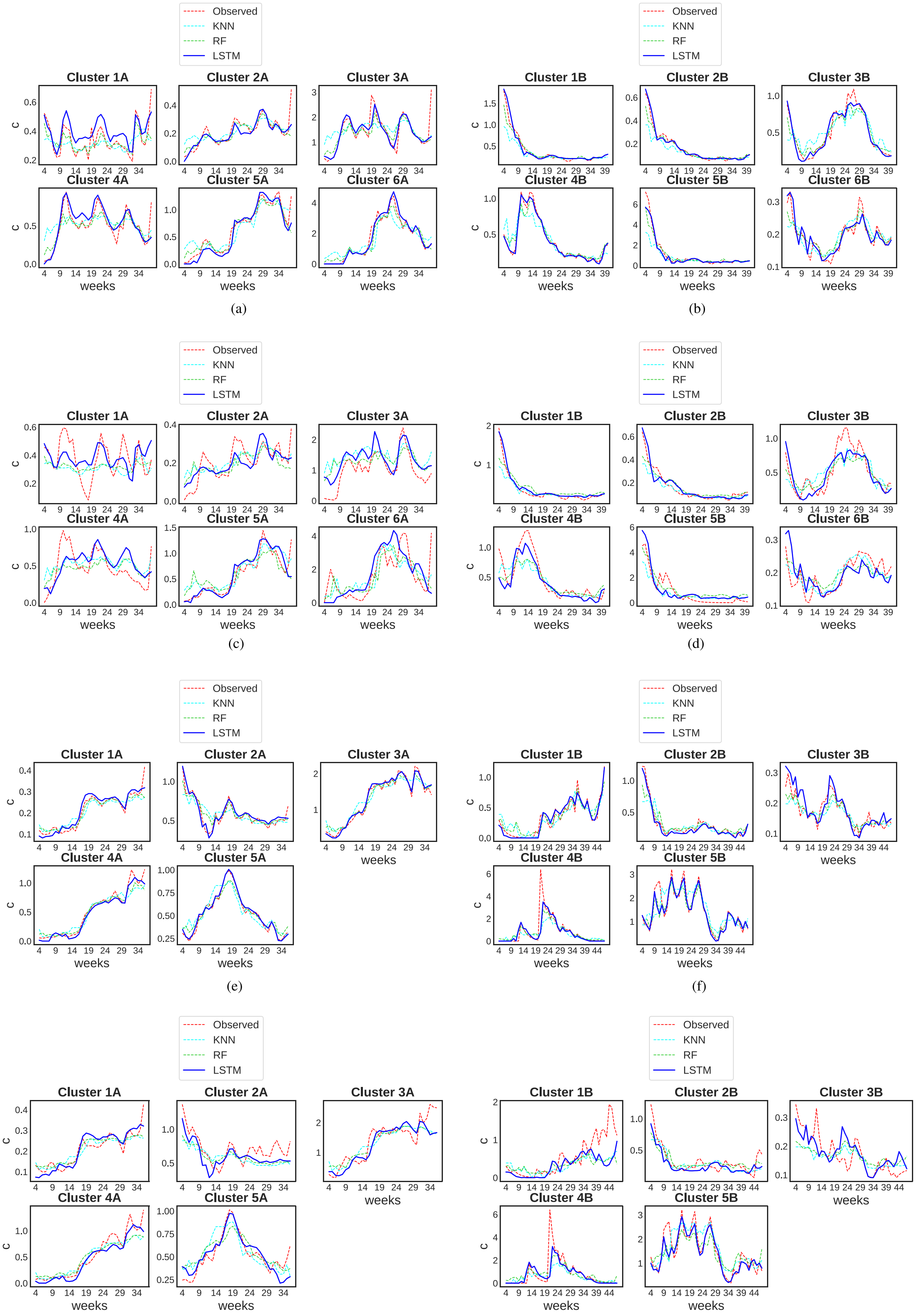
Table II presents the quality of the models resulting from the grid search for the optimal temporal window size T. Based on these data, we chose $T = 3$ as the optimal temporal window in both locations. This result is supported by [24] and [41], which shows that the development cycle of *Ae. aegypti* from egg to adult ranges between one-and-a-half to three weeks. The environmental conditions during this development period determine the transition rate of the eggs to adult. The annual best models in both locations are used in all further experiments.

Table III presents the quality of models resulting from the grid search for optimal learned representation size with T set to 3. This table shows that in Vila Velha, the learned representation size $v = 128$ produces the best quality on the test data in 2017, while $v = 16$ produces the best quality in 2018. In Serra, $v = 64$ produces the best quality on the test data in 2017, while $v = 32$ produces the best quality in 2018.

The search for learned representation size is a standard practice with fitting encoder–decoder neural network, and its result does not have a direct epidemiological bearing. However, it can be seen that the best models obtained in 2017 in both locations require higher values of v compared to their 2018 counterparts. This is because the 2017 field mosquito data contain patterns with more variability than 2018 due to an improvement of control activities (see Figs. 5 and 6).

Fig. 9 presents the line plots comparing the best LSTM models with the benchmark RF and KNN models on training (validation inclusive) and test data in both years. Fig. 10 present scatterplots comparing the models with respect to their fitness to test data. Spikes in the line plots are indicative of increasing rate of vector population. From an epidemiological standpoint, these spikes are proxies to increasing risk of diseases occurrence in neighborhoods around the cluster component trap points. Hence, forecasting such spikes will serve well as disease outbreak early warning signals. Dips in the line plots, contrarily, are indicative of low rates of vector population. The ability to forecast dips correctly in all clusters is also important, since it may lead to better resource allocation through the redeployment of control resources from areas with predicted dips to areas with predicted spikes.

The resulting best LSTM models from Table III were compared with their corresponding baseline RF and KNN models; Table IV shows the results. In Vila Velha, LSTM performs approximately 9% and 11% better than RF in 2017 and 2018, respectively. Also, it performs approximately 9% and 6% better than KNN in 2017 and 2018, respectively. It can also be seen that LSTM always perform much better than both baseline models on validation data in this location and in both years. In Serra, LSTM produces an improvements of approximately 12% and 21% with respect to RF in 2017 and 2018, respectively. In addition, it performs around 19% and 20% better than KNN in



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 9. Line plots comparing observed and predicted values for LSTM, KNN, and RF models in 2017 and 2018. Validation data points are inserted into their time positions among the training data. The weeks for each plot are indexed starting from 1 to enhance plots visualization. These weeks can be mapped to their corresponding epidemiological weeks defined using figures for the corresponding location and year in Fig. 5. (a) Vila velha, training, 2017. (b) Vila velha, training, 2018. (c) Vila velha, testing, 2017. (c) Vila velha, testing, 2017. (d) Vila velha, testing, 2018. (e) Serra, training, 2017. (f) Serra, training, 2018. (g) Serra, testing, 2017. (h) Serra, testing, 2018.

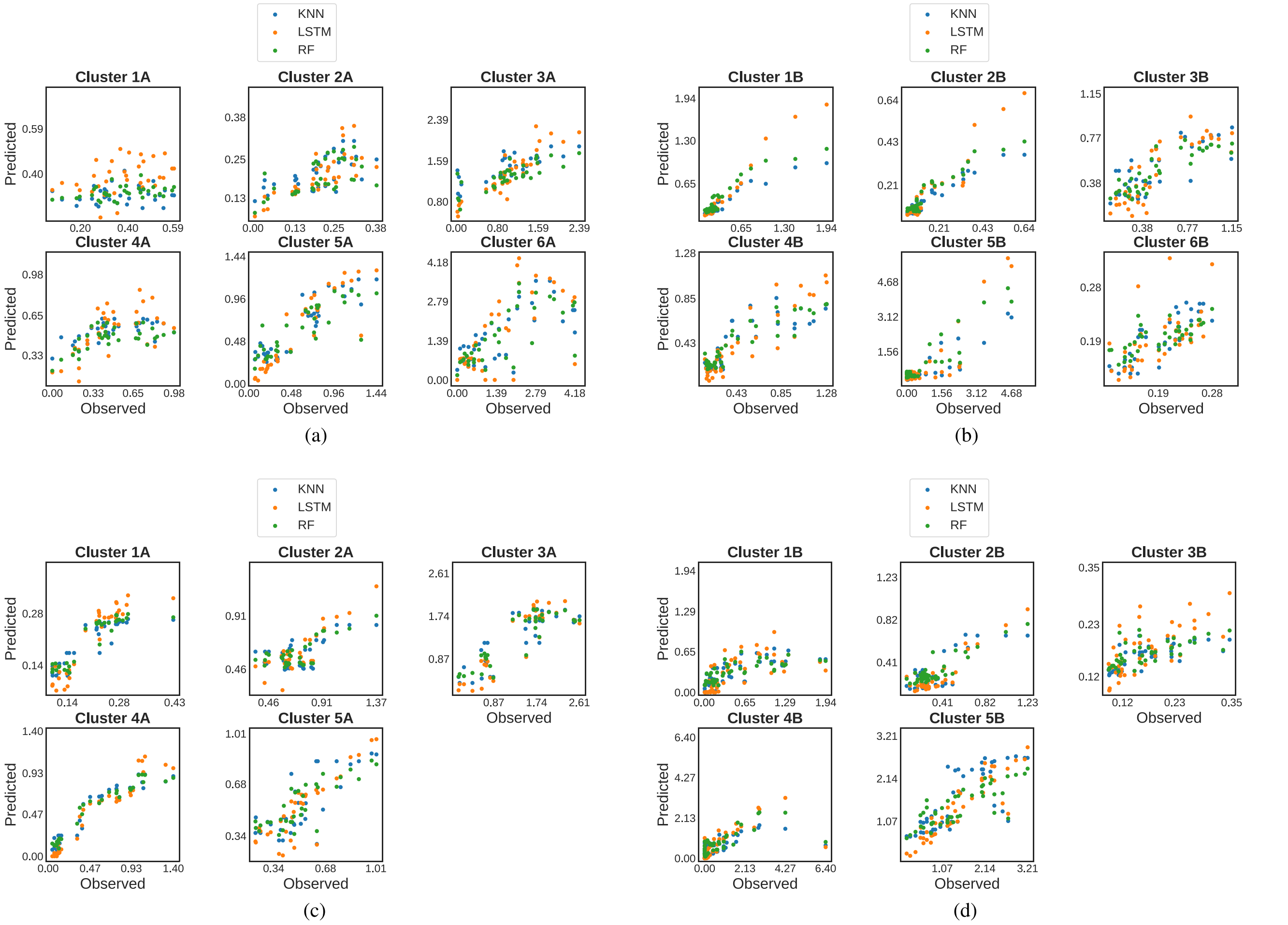


Fig. 10. Scatterplots comparing observed and predicted values for LSTM, KNN and RF models on test data. (a) Vila velha, 2017. (b) Vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

TABLE IV
Comparison of MAE for Best LSTM, RF, and KNN in Both Considered Years

		Vila Velha		Serra	
Model	Year ⇒	2017	2018	2017	2018
LSTM	Training	0.2802	0.1407	0.1630	0.2652
	Validation	0.3880	0.1998	0.1472	0.1817
	Test	**0.5767**	**0.3600**	**0.3984**	**0.3986**
RF	Training	0.1660	0.1300	0.1424	0.1743
	Validation	0.8822	0.3920	0.3940	0.4709
	Test	0.6348	0.4045	0.4502	0.5057
KNN	Training	0.4237	0.3167	0.3067	0.3643
	Validation	0.9755	0.5010	0.4165	0.6127
	Test	0.6676	0.3869	0.4906	0.5000

observation years 2017 and 2018, respectively. The validation error for LSTM is lower than what is obtained with RF and KNN in both observation years. It is worth recalling at this point that LSTMs leverage the sequential ordering of the input data in the learning process. This is especially useful for learning lagged contributions of predictor features across time. Our results here show significant quantitative evidence of the need for this property of LSTMs for the specific use case addressed in this study.

Fig. 9(a) and (b) presents the line plots of LSTM and the baseline models (RF and KNN) on training data in Vila Velha across the two observed years. LSTM mostly overestimates the observed training data in observation weeks 9–34 in Cluster 1A [see Fig. 9(a)], but still follows the observed trend. RF and KNN do not follow the observed data trend in this period. RF and KNN underestimate the observed values in the neighborhood of observation weeks 4–9 in Cluster 1A, and overestimate around these same weeks in the remaining clusters. LSTM, on the other hand, fits the data well in that period. All the fitted models fail to reach the observed data value in week 36 in Clusters 1A–5A, but LSTM significantly performs better in that week in Cluster 1A. In 2018 [see Fig. 5(b)], RF and KNN both underestimate the observed data around observation weeks 4–6 in all clusters, except in Cluster 4B. LSTM, on the other hand, fits the data well in these weeks in all the clusters, except in Cluster 5B. Also, RF and KNN underestimate the observed training data around weeks 9–14 in Cluster 4B, and overestimate their predictions in this same period in Cluster 3B.

Still on Vila Velha, with regards to test data performance, in 2017, as presented in Fig. 9(c), LSTM, KNN, and RF do not fit the observed test data well in Cluster 1A compared to the other clusters. This can be attributed to the lower purity of this cluster, as can be inferred from the differences between the training and test data temporal distribution (compare Cluster 1A training and test patterns in Fig. 9(a) and (c), respectively). Nevertheless, for this same cluster, LSTM still follows the trend (spikes and dips) of the vector population in weeks 9–14, 19–24, and 31–36, which is a total of 18 out of 36 weeks. RF and KNN, on the other hand, remain quasi-invariant in temporal pattern all through the observation weeks.

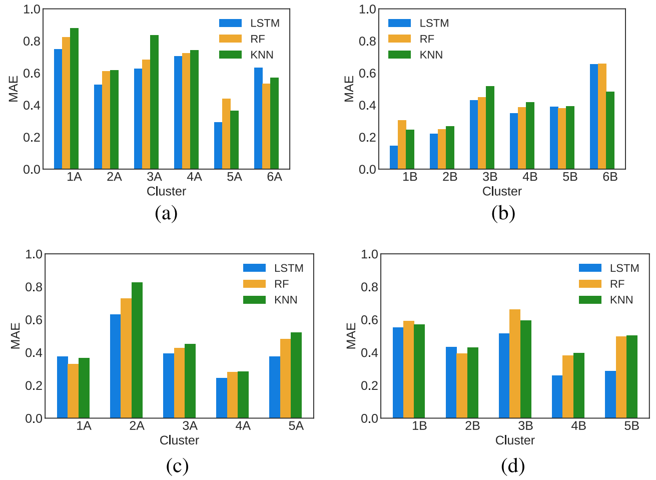
These results show the robustness of the LSTM to clustering quality variations, which is a major component of the framework proposed in this study. It is worth mentioning, however, that our test results can be improved by improving the clustering procedure. In the other clusters, around weeks 4–9, RF and KNN wrongly predict a spike in Clusters 2A–5A, while LSTM performs better in that period in the mentioned clusters. In epidemiological terms, overestimation (e.g., wrongly forecasting a spike) of vector population, as exhibited here by RF, can result in false outbreak alarms.

Considering the test data performance in Vila Velha for year 2018 as presented in Fig. 9(d), RF underestimates the observed data around weeks 4–9 in most clusters. LSTM, however, fits the observed data in all but Cluster 6B during these weeks. Another significant discrepancy between LSTM and the baseline models is around weeks 9–14 in Cluster 4B, in which KNN and RF significantly underestimate the observed data, while LSTM produces better fit. In all, LSTM shows the ability to compensate for common weaknesses shown by the two baseline models.

Fig. 10(a) and (b) compares the prediction by all fitted models to the observed test data with a scatterplot visualization. Here, it is seen that LSTM follows the highest observed values better for all clusters in both years. Again, this is indicative for better capability to forecast possible disease outbreaks. LSTM also follows the lowest observed values better in Clusters 2A–5 A in 2017. In 2018, LSTM follows the lowest observed values better in Clusters 3B and 4B.

In Serra, first we discuss how the models perform on the training data in both years as presented in Fig. 9(e) and (f). In 2017, on the training data [see Fig. 9(e)], both RF and KNN overestimate the observed data in weeks 4–9 for all clusters except Cluster 2A, where KNN underestimates the observed data. Also, both baseline models underestimate the peak reached around observation week 19 in Cluster 5A. On the contrary, LSTM performs relatively well in all these periods. In 2018 [see Fig. 9(f)], again, both RF and KNN underestimate the observed data around the observation weeks 4–9 in clusters 2B and 3B.

With regards to the test data performance of the models in Serra, the same disparities from the observed training data shown by the baseline models in 2017 are also reproduced on the test data [see weeks 4–9 in Fig. 10(g)]. In 2018 [see Fig. 10(h)], notable disparities between observed and fitted data by all models are seen in Cluster 1B starting from observation week 39. This is another case of high intracluster variance, which has led to different patterns in training and test data of the same cluster. Regardless, for this cluster and in this period, LSTM still attempts to capture some of the temporal variations, while RF and KNN remain relatively invariant.



(a)
(b)
(c)
(d)

Fig. 11. Cluster-level comparison of MAE for LSTM, KNN and RF models. Lower MAE is desirable. (a) Vila velha, 2017. (b) Vila velha, 2018. (c) Serra, 2017. (d) Serra, 2018.

From the visualization offered by the scatterplots presented in Fig. 10(a) and (b), we observe that, just like in Vila Velha, LSTM follows the lowers observed values better in Serra for all clusters in both years. Also, LSTM follows the highest observed values better in Clusters 2A, 4A, 5A, 1B, 2B, 3B, and 5B.

Overall, LSTM qualitatively outperforms both baseline models in generalizing to the test data in both test locations. By following the highest and lower observed data better, LSTM provides the most reliable model for an outbreak early warning system.

Finally, the errors produced by the LSTM models in each location are examined at cluster level in comparison to their RF and KNN counterparts. This analysis is presented in Fig. 11. Here, it is shown that, in Vila Velha, LSTM produced the lowest MAE in five and four cluster out of the total six in 2017 and 2018, respectively. In Serra, LSTM produced better results in four out of the five clusters in both years.

VII. Discussion

According to Fig. 5, the clustering approach applied in this study succeeds in finding common patterns in the trap points series. In this way, the problem of forecasting the many underlying series obtained from the traps is simplified to that of forecasting fewer series. The pattern similarity among traps series in the same cluster is captured by the similarity in the observed training and test data, which have been obtained as averages of randomly selected series as shown in Fig. 9. By this method, we have reduced the forecasting task significantly.

In Vila Velha, starting from 195 and 325 trap points series in 2017 and 2018, respectively, we obtain six clusters for each year that describe the underlying mosquito vector activity of interest during the time observed. In Serra, starting from 567 and 95 traps in 2017 and 2018, respectively, we are able to summarize them into five clusters in each year. It is noteworthy that the optimal number of underlying clusters obtained is the same for each location in the two observation years. This shows that, in spite of different control conditions and nonmatching climate seasons in both years, the underlying pattern mechanisms of the female *Ae. aegypti* in these locations are continuous.

The results show that freely accessible satellite image products, which have formed the basis of recent studies [1], [2] in *Ae. aegypti* population dynamics modeling are available at spatial resolutions that make them informative for neighborhood-level temporal modeling. This is useful for municipality, regional, or national monitoring, where a larval survey approach is used to plan preventive and recovery actions. This approach requires that designated field inspectors visit all traps weekly at specified times to inspect and collect the data, resulting sometimes in missing data due to insufficient manpower. This issue is observable in the data of this research because, as mentioned above, among 791 traps in Vila Velha, only 193 traps had significant records for the whole year 2017. Since the cost of collecting *in situ* data is very high, the financial inefficiency resulting from of large amounts of missing data is also very high. As a result, the framework proposed in this work can serve not only for forecasting purposes, but also for spatio-temporal gap filling, especially when a trap location with missing data had previously been classified into a cluster.

Another point worth discussing is the importance of the lagging effects on some variables. Indeed, many studies have reported varying lagged effects associations between environmental conditions and dengue virus spread dynamics. In [2], two weeks of lag was chosen to represent nonsynchronous environmental effects. Scavuzzo *et al.*[1] and German *et al.* [5], on the other hand, chose three weeks for same take but in a different study location from [2]. These studies base their choice of lag window on *a priori* entomological knowledge of mosquito development life cycle. However, as reported in [42], this prior based lagged effects knowledge does not generalize globally, and is not necessarily the same for every considered environmental condition. For example, increase in dengue risk has been associated with increasing minimum and maximum temperature by 1–2 two month lags in Mexico, French, West Indies, and Brazil. Countries closer to the equator, e.g., Singapore and Indonesia, report shorter lag effects (2–4 weeks) of temperature on the dengue cases. The study in [43] presents the temporal analysis of the relationship between dengue virus (not vector) and climatic variables in Rio de Janeiro, Brazil between the years 2001 and 2009. The best result in that study was obtained by considering four weeks ($T = 4$) lag effects of both precipitation and temperature variables. In line with all these works, our results show that $T = 3$ is the most significant choice, in accordance with empirical evidence.

In our study, we have considered an experimental approach toward choosing the right temporal window not only in terms of size, but maintaining the sequential ordering of the considered lagged series. Indeed, a major advantage of RNNs is that they can, within a specified time window, automatically learn the right lag dependencies differently for each considered environmental variable feature. As shown by the results in Section VI-D, learning the sequential dependency in lagged temporal windows improves the quality of our model. This improvement in quality generalizes across multiple vector control regimes and in two different locations.

As already mentioned in Section VI-E, one way to improve the presented results is by improving the quality of the clusters obtained. In this work, we have chosen to cluster the mosquito data based solely on their time series characteristics. This approach can be improved by also considering geospatial proximity of trap locations.

VIII. Conclusion

While in [2], an RF approach was exploited for municipality-level "nowcasting" of *Ae. aegypti* vector population, in this work, the same satellite image features were used to design a neighborhood-level forecasting framework. To this aim, autoregressive (past vector dynamics) and exogeneous (environmental effects) components were both included in the proposed model, and RNN were used to learn the model parameters and sequential dependency, especially considering lagged effects.

Eventually, this study results in the following contributions.

1) A general RNN-based algorithm for neighborhood-level time series female *Ae. aegypti* population one-week-ahead forecasting using EO products has been proposed and validated.
2) Forecasting accuracy values better than those by multi-output variants of RF (as applied in [2]) and k-Nearest Neighbor (KNN) (as applied in [1]) have been obtained.
3) By applying our modeling pipeline to data from different time periods and locations, the proposed approach has been proved as robust and with generalization capabilities for different conditions.
4) Finally, we have discussed how the results obtained from the proposed method can be applied to improve existing vector surveillance systems in terms of cost, time, and man-power efficiency.

Using the one-week-ahead NARX forecast model proposed in this work, public health managers have more time to plan and respond. Future studies will consider an extension into multisteps-ahead forecasting, since RNNs have already been successfully applied to such cases in other domains [17], [44]. Additionally, other deep learning network architectures will be considered, such as the transformer model applied in [9] for multistep-ahead forecasting of influenza epidemics prevalence. Another equally promising future direction is toward using the resulting RNN models to understand the relative importance of the observed environmental factors on the population of female *Ae. aegypti*. This is particularly challenging using deep learning techniques, as testified by the growing number of researches on the topics of "explicable AI" and "whitening AI."

Acknowledgment

The authors would like to thank the field workers from Vila Velha and Serra for collecting field data, the health managers of these municipalities: Marcelaine Marculano and Gilberto Santos, and the Espírito Santo Health Department for granting them permission to use the mosquito population data.

References

[1] J. M. Scavuzzo *et al.*, "Modeling dengue vector population using remotely sensed data and machine learning," *Acta Tropica*, vol. 185, pp. 167–175, Sep. 2018.

[2] O. Mudele, F. M. Bayer, L. F. Zanandrez, A. E. Eiras, and P. Gamba, "Modeling the temporal population distribution of *Ae. aegypti* mosquito using big earth observation data," *IEEE Access*, vol. 8, pp. 14 182–14 194, 2020.

[3] J. P. Messina *et al.*, "Mapping global environmental suitability for Zika virus,*" Elife*, vol. 5, 2016, Art. no. e 15272.

[4] D. P. O. de Melo, L. R. Scherrer, and A. E. Eiras, "Dengue fever occurrence and vector detection by larval survey, ovitrap and mosquitrap: A space-time clusters analysis," *PLoS One*, vol. 7, no. 7, p. e42125, 2012.

[5] A. German, M. Espinosa, M. Abril, and C. Scavuzzo, "Exploring satellite based temporal forecast modelling of Aedes Aegypti oviposition from an operational perspective," *Remote Sens. Appl., Soc. Environ.*, vol. 11, pp. 231–240, 2018.

[6] E. Parselia *et al.*, "Satellite earth observation data in epidemiological modeling of malaria, dengue and west nile virus: A scoping review," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1862.

[7] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA, Princeton university press, 1994, vol. 2.

[8] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, "Analog forecasting of extreme-causing weather patterns using deep learning," *J. Adv. Model. Earth Syst.*, vol. 12, no. 2, p. e2019MS001958, 2020.

[9] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," 2020, *arXiv:2001.08317*.

[10] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, AAAI Press, 2017, pp. 2627–2633.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[12] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Mach. Learning*, vol. 7, no. 2-3, pp. 195–225, 1991.

[13] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," *Wseas Trans. Comput. Res.*, vol. 3, no. 3, pp. 182–191, 2008.

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018.

[15] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 473–479.

[16] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. 2014 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pp. 1724–1734, 2014.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[18] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*. 2013, pp. 6645–6649.

[19] V. Andreo, X. Porcasi, C. Rodriguez, L. Lopez, C. Guzman, and C. M. Scavuzzo, "Time series clustering applied to eco-epidemiology: The case of Aedes Aegypti in Córdoba, Argentina," in *Proc. XVIII Workshop Inf. Process. Control*, 2019, pp. 93–98.

[20] O. J. Brady *et al.*, "Modelling adult Aedes Aegypti and Aedes Albopictus survival at different temperatures in laboratory and field settings," *Parasites Vectors*, vol. 6, no. 1, 2013, Art. no. 351.

[21] S. Chandy *et al.*, "Assessing effect of climate on the incidence of dengue in tamil nadu," *Indian J. Med. Microbiol.*, vol. 31, no. 3, 2013, Art. no. 283.

[22] K. M. Campbell *et al.*, "Weather regulates location, timing, and intensity of dengue virus transmission between humans and mosquitoes," *PLoS Neglected Trop. Dis.*, vol. 9, no. 7, p. e0003957, 2015.

[23] S. Sang *et al.*, "Predicting local dengue transmission in Guangzhou, China, through the influence of imported cases, mosquito density and climate variability," *PLoS One*, vol. 9, no. 7, p. e102755, 2014.

[24] A. Hussain, F. Ali, O. B. Latiwesh, and S. Hussain, "A comprehensive review of the manifestations and pathogenesis of Zika virus in neonates and adults," *Cureus*, vol. 10, no. 9, p. e3290, 2018.

[25] D. W. Allgood and D. A. Yee, "Oviposition preference and offspring performance in container breeding mosquitoes: Evaluating the effects of organic compounds and laboratory colonisation," *Ecol. Entomol.*, vol. 42, no. 4, pp. 506–516, 2017.

[26] B. C. Gao, "NDWIa normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 257–266, Dec. 1996.

[27] J. Xue and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," *J. Sensors*, vol. 2017, pp. 1–17, 2017.

[28] Z. Jiang, A. R. Huete, K. Didan, and T. Miura, "Development of a two-band enhanced vegetation index without a blue band," *Remote Sensing Environ.*, vol. 112, no. 10, pp. 3833–3845, 2008.

[29] G. Tang, M. P. Clark, S. M. Papalexiou, Z. Ma, and Y. Hong, "Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM imerg with nine satellite and reanalysis datasets," *Remote Sens. Environ.*, vol. 240, 2020, Art. no. 111697.

[30] G. Skofronick-Jackson *et al.*, "The global precipitation measurement (GPM) mission for science and society," *Bull. Amer. Meteorol. Soc.*, vol. 98, no. 8, pp. 1679–1695, 2017.

[31] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi, "Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 1672–1678.

[32] X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, and S. Yan, "Deep learning with s-shaped rectified linear activation units," in *Proc. 13th AAAI Conf. Artif. Intell.*, vol. 30. no. 1, 2016.

[33] A. Kassambara, *Practical guide to cluster analysis in R*, 1st ed. New York, N.Y. USA: STHDA, 2017.

[34] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1855–1870.

[35] A. Ng, "Clustering with the *k*-means algorithm," *Mach. Learn.*, 2012.

[36] Brasil. Ministério da Saúde, "Boletim Epidemiológico," 2020. Accessed: Oct. 5, 2020. [Online]. Available: https://antigo.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf

[37] K. M. Pepin, C. Marques-Toledo, L. Scherer, M. M. Morais, B. Ellis, and A. E. Eiras, "Cost-effectiveness of novel system of mosquito surveillance and control, Brazil," *Emerg. Infect. Dis.*, vol. 19, no. 4, 2013, Art. no. 542.

[38] Z. Wan, Y. Zhang, Q. Zhang, and Z.-L. Li, "Quality assessment and validation of the MODIS global land surface temperature," *Int. J. Remote Sens.*, vol. 25, no. 1, pp. 261–274, Jan. 2004.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Mach. Learn. Appl., Int. J.*, vol. 3, no. 2, pp. 19–28, 2016.

[41] V. S. H. Rao and R. Durvasula, *Dynamic Models of Infectious Diseases*. New York, NY, USA: Springer, 2013, vol. 1.

[42] Y. L. Cheong, K. Burkart, P. J. Leitão, and T. Lakes, "Assessing weather effects on dengue disease in Malaysia," *Int. J. Environ. Res. Public Health*, vol. 10, no. 12, pp. 6319–6334, 2013.

[43] A. F. Gomes, A. A. Nobre, and O. G. Cruz, "Temporal analysis of the relationship between dengue and meteorological variables in the city of Rio de Janeiro, Brazil, 2001–2009," *Cadernos de Saúde Pública*, vol. 28, no. 11, pp. 2189–2197, 2012.

[44] I. M. Kamal, H. Bae, S. Sunghyun, and H. Yun, "DERN: Deep ensemble learning model for short-and long-term prediction of baltic dry index," *Appl. Sci.*, vol. 10, no. 4, 2020, Art. no. 1504.

Oladimeji Mudele received the bachelor of engineering degree in electrical and electronic engineering from the Federal University of Technology, Akure, Nigeria, in 2013, and the master's degree (*cum laude*) in electronic engineering from The University of Pavia, Pavia, Italy, where he is currently working toward the Ph.D. degree with the Telecommunications and Remote Sensing Laboratory.

Mr. Mudele was the recipient of the Premio di Laurea Award from the Rotary International Club, Pavia, Italy, for excellent contributions as an international student in The University of Pavia, in February 2019.

Alejandro C. Frery (Senior Member, IEEE) received the B.Sc. degree in electronic and electrical engineering from the Universidad de Mendoza, Mendoza, Argentina and the M.Sc. degree in applied mathematics (Statistics) from the Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil, and the Ph.D. degree in applied computing from the Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil.

He is currently a Professor in statistics and data science in the Victoria University of Wellington, New Zealand. His research interests are statistical computing and stochastic modeling.

Prof. Frery was the founder of LaCCAN – *Laboratório de Computação Científica e Análise Numérica*, Universidade Federal de Alagoas, Maceió, Brazil, and holds a Huashan Scholar position (2019–2021) with the Key Lab of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, Xi'an, China. He was the Editor-in-Chief of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2014–2018). In 2018, he was the recipient of the IEEE GRSS Regional Leader Award.

Lucas F. R Zanandrez received the B.S. degree in biomedicine from the Federal University of Minas Gerais, Brazil, in 2016, with an exchange period at Monash University in Melbourne, Australia and the M.S. degree in technological innovation and intellectual property from the Federal University of Minas Gerais in 2019.

His current research interests include the study of infectious diseases and the development of innovations that can be applied to public health.

Alvaro E. Eiras received the Ph.D. degree in chemical ecology from the University of Southampton, Southampton, U.K., in 1991, the Postdoctoral degrees from the University of Viçosa, MG, Brazil and James Cook University, Queensland, Australia, in 1993, and 2013, respectively.

He has authored or coauthored about 100 articles in international peer-review journals and presented more than 200 research works in workshops and conferences. He is currently a Full Professor at the Federal University of Minas Gerais (UFMG), where he leads the Laboratory of Technological Innovation and Entrepreneurship in Vector Control, which developed innovated mosquito traps and synthetics attractants and, an Aedes surveillance system at real time.

Dr. Eiras was a recipient of the two international awards for Technological Innovation: Tech Award Museum (San Jose, Silicon Valley, CA, USA), and the Edison Awards - Innovations & Innovators (Chicago, IL, USA).

Paolo Gamba (Fellow, IEEE) received the Laurea degree "cum laude" and the Ph.D. degree in electronic engineering from the University of Pavia, Pavia, Italy, in 1989 and 1993, respectively.

He is currently a Professor with the University of Pavia, Pavia, Italy, where he leads the Telecommunications and Remote Sensing Laboratory. He has authored or coauthored more than 170 papers in international peer-review journals and presented 310 research works in workshops and conferences.

Dr. Gamba served as an Editor-in-Chief of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2009 to 2013, and as Chair of the Data Fusion Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) from October 2005 to May 2009. He has been elected in the GRSS AdCom since 2014, served as GRSS President from 2019 to 2020, and is currently GRSS Junior Past President. He has been the Organizer and Technical Chair of the biennial GRSS/ISPRS Joint Workshops on "Remote Sensing and Data Fusion over Urban Areas" from 2001 to 2015. He also served as Technical Co-Chair of the 2010, 2015, and 2020 IGARSS conferences, in Honolulu (Hawaii), Milan (Italy), and online, respectively. He has been invited to give keynote lectures and tutorials in several occasions about urban remote sensing, data fusion, EO data for physical exposure, and risk management.