IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 14, 2021, accepted October 4, 2021, date of publication October 14, 2021, date of current version October 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120309

A Windowed Correlation-Based Feature Selection Method to Improve Time Series Prediction of Dengue Fever Cases

TANVIR FERDOUSI[1], **(Member, IEEE), LEE W. COHNSTAEDT**[2], **AND CATERINA M. SCOGLIO**[1], **(Senior Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA
[2]Foreign Arthropod-Borne Animal Diseases Research Unit, National Bio and Agro-Defense Facility, United States Department of Agriculture (USDA)-ARS, Manhattan, KS 66502, USA

Corresponding author: Tanvir Ferdousi (tanvirf@ksu.edu)

This work was supported by the Department of the Army, U.S. Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Fort Detrick, under DWFP Grant W911QY-19-1-0004.

ABSTRACT The performance of data-driven models depends on training samples. For accurately predicting dengue fever cases, historical incidence data are inadequate in many locations. This work aims to enhance temporally limited dengue case data by methodological addition of epidemically relevant case data from nearby locations as predictors (features). A novel framework is presented for windowing incidence data and computing time-shifted correlation-based metrics to quantify feature relevance. The framework ranks incidence data of adjacent locations around a target by combining metrics based on correlation, spatial distance, and local prevalence. Recurrent neural network models achieve up to 33.6% accuracy improvement on average using the proposed method. These models achieve mean absolute error (MAE) values as low as 0.128 on [0,1] normalized incidence data for a municipality with the highest dengue prevalence in Brazil's Espirito Santo. When predicting aggregate cases over geographical ecoregions, the models improve by 16.5%, using only 6.5% of ranked incidence data. This paper also presents two correlation window allocation methods: fixed-size and outbreak detection. Both perform comparably well, although the outbreak detection method uses less data for computations. The proposed framework is generalized, and it can be used to improve time-series predictions of many spatiotemporal datasets.

INDEX TERMS Dengue, feature selection, machine learning, recurrent neural networks, time series.

I. INTRODUCTION

Accurate time series prediction of dengue fever outbreaks can be useful in planning mitigation strategies for hundreds of tropical and subtropical regions around the world. Data-driven models such as neural networks are flexible in design and can ease the difficulties of estimating unknown parameters that mechanistic models often require [1]. However, the prediction accuracy of such models depends on the quality and the quantity of training data. For dengue fever outbreaks, the availability of incidence data varies across regions. Thus, a data aggregation center in an area may not have adequate data to achieve an acceptable level of accuracy in out-of-sample projections. In such cases, selected incidence data from adjacent centers in the same region could improve model performance as additional features. We propose a framework with quantitative methodologies to rank and select nearby case data as supplementary features. Our method uses windowed time-lagged cross-correlation combined with distance and prevalence metrics to identify relevances and potential causal relationships.

Dengue virus is primarily spread by several species of mosquito vectors (*Aedes aegypti* and *Aedes albopictus*), and the outbreaks infect 390 million people every year [2]. The viral strains also cause about 40,000 annual deaths with hemorrhagic fever and dengue shock syndrome [3]. Dengue virus transmission is prevalent in regions where competent vector mosquitoes are present. In those regions, the mosquito abundance varies throughout the year and depends on factors including air temperature, precipitation,

The associate editor coordinating the review of this manuscript and approving it for publication was Kaustubh Raosaheb Patil.

vegetation, and urbanization [4]–[7]. However, researchers experience missing data, uneven reporting intervals, lack of granularity, inadequacy, and inaccuracy with dengue case counts for many regions around the globe. For most diseases, time-series outbreak data are non-stationary (i.e., the underlying processes evolve with time). Outbreak start times and sizes can vary because of imported cases caused by short and long-range mobility. Co-circulation of multiple viral strains adds to the complexity by changing immunity patterns in host populations. Despite those issues, correlations exist between outbreaks in adjacent human populations (county, municipality, district, etc.) for most infectious diseases, including dengue [8]. While a correlation may not always imply causation, incidence data from adjacent regions can improve the model training because of similarities in meteorological factors, host population density, and a high probability of mobility-based viral transmissions.

For sequential data (e.g., multivariate time series), different methods of feature selection have been used in the past including correlation-based filters [9], [10], Granger causality tests [11], [12], genetic algorithm [13], wavelet transform with nested long short-term memory networks [14], and several other methods [15], [16]. Only a few works extend feature sets for disease outbreak prediction using incidence data from spatially adjacent locations. Such geospatial clustering techniques rely on similarity metrics to improve model performance. A recent work [17] uses pair-wise correlation to cluster location data to train models for COVID-19 outbreak projection in Chinese provinces. Another work targeted towards dengue also uses correlation-based similarity measures to extend the training feature set [18]. However, these implementations assume an instantaneous correlation of incidence data between regions and do not consider the temporal order of outbreaks. There can be a considerable amount of time delay between outbreaks occurring in adjacent regions. Time lagged cross-correlation can help identify such phase relations [19]. The phase information can quantify relationships between outbreaks of adjacent locations. We hypothesize that *if one outbreak leads another with respect to time, the former outbreak may have a causal influence on the latter*, given that these places lie spatially close enough to affect one another, and outbreak sizes exceed a reasonable threshold. Based on this idea, we rank incidence data using a combination of these factors: leading phase correlation, distance, and prevalence. Because of dynamically changing immunity patterns, outbreak hot spots can also evolve. A single computation of correlation may mislead the analysis. Hence, we split each time series into multiple windows. Previous works have used different ways to segment temporal data. For example, recurrence plot-based analysis has been used to segment multivariate time series [20]. However, the evolving dynamics of dengue viral strains and host immunity patterns have a complex relationship with outbreak start/end times. Hence, a splitting method tailored to dengue outbreaks is necessary.

This work presents a novel framework of feature selection for the data-driven prediction of dengue outbreaks. This framework ranks and selects incidence data from spatially adjacent locations with significant prevalence and leading positive correlation. To achieve that goal, this paper details: i) a windowed time-shifted cross-correlation method to compute correlation weights, ii) two window allocation methods to segment time series data, iii) a procedure of ranking feature using a combination of correlation, distance, and prevalence metrics, and iv) analysis of prediction performance across windowing methods, prediction models, and spatial aggregations. The originality of this work lies in how we interpret and process incidence data to compute correlation metrics using our knowledge of how outbreaks spread.

II. PRELIMINARIES

A. DEFINITIONS

In supervised learning problems, we collect data on multiple variables. A *target* variable (*label*) is the designated output of a machine learning model for prediction. A *feature* is an input variable that is expected to influence the target variable. Each *instance* of data (i.e., a point in time) contains several feature values and usually a single label value. A data set comprises many such instances. For a supervised learning problem, a data set is split into 3 subsets in order to: *train*, *evaluate*, and *test* the models. With the training subset fed in batches (collection of instances), a supervised model learns to predict targets based on features in an iterative process. It optimizes parameters by minimizing a loss function. A neural network comprises artificial neurons (cells) in one or multiple layers. Each neural cell is a node in the network with connections to other nodes across layers, inputs, or outputs. The recurrent neural networks (RNN) are special neural cells that store internal states in their memory, which helps them predict sequence data. Model training aims to get optimal parameter values to generalize beyond the training data and perform well with test data. Sometimes, a model can over-fit the training data and perform poorly with unseen test examples. To prevent such scenarios, we use regularization techniques.

B. TIME SERIES PREDICTION OF OUTBREAKS

In epidemiology, time series models enable the prediction of future outbreaks by fitting models with past disease incidence data and carefully chosen covariates. Such predictions come at varying degrees of accuracy and depend on the characteristics of the target variable, quality of sample data used for fitting, covariates, and the models themselves. Classical models such as exponential smoothing, ARIMA, and seasonal ARIMA tend to follow mean values of past data [21], [22], and it is not easy to associate them with rapidly changing processes [23]. In addition to that, these models require manual tuning of their parameters and may fail to capture complex nonlinear interactions. Data-driven forecasting of vector-borne diseases such as dengue fever is difficult due to the complex interactions of several factors with disease dynamics. A list of these factors include but is not limited to seasonally dependent *Aedes* mosquito growth and feeding patterns [24], co-circulations of multiple viral strains [25], environmental (e.g., temperature) effects on dengue virus transmission [26], and human mobility patterns [27]. Neural networks can automatically interpret features from observable variables and learn complex nonlinear phenomena. Long-short term memory (LSTM) network [28] (a type of recurrent neural network), and its derivatives [29], [30] have shown promise in predicting sequential data. These have frequently outperformed classical models [31]. Recurrent neural architectures have also performed well in forecasting disease outbreaks [32]–[34]. Hence, we consider these as viable candidates to test the performance of our proposed feature enhancement framework.

C. FACTORS THAT AFFECT DENGUE DISEASE DYNAMICS

Dengue virus is primarily spread by female mosquito vector species: (*Aedes aegypti* and *Aedes albopictus*). Hence, the outbreaks depend on the abundance of such vectors. The relationships between *Aedes* mosquitoes and environmental variables (e.g., temperature, rainfall) are already well characterized by many researchers. Environmental temperature affects the growth, host-seeking, blood-meal intakes of mosquitoes. *Aedes aegypti* cannot develop below 16° C or above 34° C [35]. Within that range, the development from larva to adult was found to be faster at higher temperatures (30° C) compared to lower temperatures (21° C) [36]. *Aedes albopictus* can develop in wider temperature ranges and can survive better in lower temperatures [37] compared to *Aedes aegypti*. The optimum flight temperature for *Aedes aegypti* females was found to be 21° C [38]. Studies have observed that a large diurnal temperature range decreases female fecundity [39]. Temperature fluctuations also affect extrinsic incubation periods (EIP) of dengue viruses. An experiment with DEN-2 strain found that EIP was 12 days at 30° C and reduced to 7 days for 32° C and 35° C [40]. Besides temperature, rainfall has a significant role in dengue outbreaks. Rainwater stuck in different places creates breeding spaces for *Aedes* mosquitoes. Previous works have studied the association of dengue transmission with rainfall [4], [41]. In this work, we use several variables, including observed and reanalyzed temperatures, precipitations, relative humidity, and surface-level pressure. These can directly or indirectly affect mosquito vector suitability and dengue outbreak dynamics.

D. DATA COLLECTION AND PROCESSING

1) DATA ACQUISITION

To test the proposed framework for dengue outbreak predictions, we collect data for several regions of Brazil. The InfoDengue project [42] monitors outbreak data on over 700 municipalities of Brazil. Their server contains weekly dengue fever case counts with a surveillance period starting from 2010. Besides dengue incidence data, we collect weather observation data from NOAA (National Oceanic and Atmospheric Administration) ground weather station database and reanalysis data from the NCEP /NCAR Reanalysis 1 dataset published by the NOAA physical sciences laboratory (PSL) database (NCEP and NCAR stand for National Centers for Environmental Prediction and National Center for Atmospheric Research, respectively). We list all the variables in Table 1. We use the weekly case counts as labels and further process the remaining variables to use those as baseline features.

TABLE 1. Data collection sources.

Variable Name	Time Resolution	Space Resolution
Dengue fever cases [a]	Weekly	Municipality level
Observed Temperatures [b]	Daily	Ground station dependent
Observed Precipitation [b]	Daily	Ground station dependent
Avg surface air temperature [c]	Daily	2.5° × 2.5°
Avg surface relative humidity [c]	Daily	2.5° × 2.5°
Avg surface pressure [c]	Daily	2.5° × 2.5°
Avg precipitable water [c]	Daily	2.5° × 2.5°

[a] InfoDengue Project [42]
[b] NCEI-NOAA ground station data
[c] NCEP/NCAR Reanalysis 1 dataset

2) RE-SAMPLING AND FEATURE ENGINEERING

The available raw data cannot be readily used in the machine learning methods. All the features and labels are matched and aligned in both spatial and temporal dimensions. We align data based on available labels (i.e., case data). For each municipality of Brazil (smallest spatial unit available), we search for the nearest ground weather station to collect weather data. We extract reanalysis data from NCEP/NCAR Reanalysis 1 data sets using each municipality's centroid's coordinates. Once the spatial granularity is taken care of, we fix the temporal dimension mismatches by converting all data to match incidence data resolution. As incidence data are available in weekly intervals and the remaining variables are available daily, this process involves context-aware down-sampling of those remaining variables (temperature, precipitation, humidity, etc.). During this process, we derive 4 additional features from observed weather data of ground stations: average diurnal temperature range of the week, minimum diurnal temperature range of the week, maximum diurnal temperature range of the week, and the number of rainy days in the week. We compute these from daily observed temperature and precipitation data. In total, we have 12 feature variables related to weather and environment: i) 4 observed variables: temperature (average, minimum, and maximum) and precipitation, ii) 4 derived variables based on the time interval (week): diurnal temperature range (average, minimum, and maximum) and the number of rainy days, iii) 4 reanalysis variables: temperature, relative humidity, pressure, precipitable water (all are averages and at earth surface level).

3) DATA SPLITTING AND NORMALIZATION

The complete set of data is a two-dimensional array X with *features* on one dimension and *time* on the other.

The set of features consists of: i) the variables listed in section II-D2, ii) dengue incidence data of the target location, and iii) dengue incidence data of locations selected as predictors by the methods presented in this paper. We split X along its time dimension into three parts: training (X_{train}), validation (X_{val}), and test (X_{test}). The validation set is required during the training phase as we implement *early stopping* [43] as a regularization mechanism. We normalize all three sets of data before model training and evaluation. The normalization formulas are given in (1).

$$\begin{aligned}\mu_{train} &= MEAN(X_{train})\\ \sigma_{train} &= SD(X_{train})\\ \hat{X_{train}} &= (X_{train} - \mu_{train})/\sigma_{train}\\ \hat{X_{val}} &= (X_{val} - \mu_{train})/\sigma_{train}\\ \hat{X_{test}} &= (X_{test} - \mu_{train})/\sigma_{train} \qquad (1)\end{aligned}$$

All three data sets (training, validation, and test) are normalized using the mean and the standard deviation computed from training data. The complete data set's summary statistics are not used here to prevent the machine learning models from gaining statistical insights about validation and test data sets.

E. SEQUENCE MODEL SPECIFICATIONS

Whether it is training or evaluation, the time series data are split into smaller batches and fed into the recurrent neural network models. From a macroscopic perspective, a sliding window moves over batches of data. Because of the sequential nature of dengue case data, the batches are fed according to the time order without randomization. The sliding window is configured with two integer parameters: input length (t_{in}), and output length (t_{out}). The window is depicted in Figure 1. A trainable model would take t_{in} time steps of feature data as input and predict t_{out} time steps of target/label data (e.g., dengue case counts) as outputs every iteration. In the configurations used in this paper, there is no temporal overlap between input and output sequences. In this configuration, the models make single shot projections (all the t_{out} data points are predicted at once every iteration).

Each batch (window) of data is $(t_{in} + t_{out})$ steps long in the time dimension. A total of 32 batches are stacked together for model training and evaluation. One batch differs from another by a single time-step (hence, there are temporal overlaps between batches). Each batch is further split in time and data (variable) dimensions to separate inputs (t_{in} of features) and outputs (t_{out} of labels).

This work focuses on performance gains obtainable using recurrent neural network (RNN) models due to their proven track record in predicting time series data [31]. A recurrent unit's temporal behavior is illustrated in the lower part of Figure 1. We can see that information is passed through time (also called cell state), enabling the model to predict values based on insights gained from past inputs. We use two popular recurrent neural network cell types: LSTM (long-short term memory) [28] with forget gates [44] and GRU (gated recurrent unit) [45]. We also test with a simple linear neural network model as a trivial baseline. Using the TensorFlow [46] package, we configure the models as described below to produce the results:

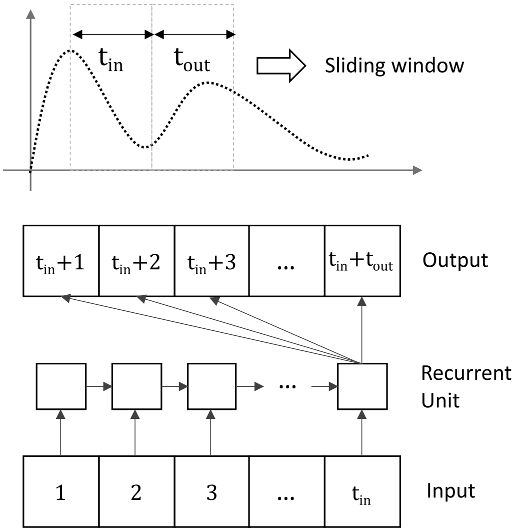

FIGURE 1. A sliding window defined for feeding of input data and extraction of output case counts. The window slides along the horizontal axis and feeds t_{in} time steps of feature data into the model and extracts t_{out} time steps of predictions.

- *Linear:* The model consists of a single layer of artificial neurons (*Dense* units in TensorFlow) without any nonlinear activation functions. The size of the layer (number of neural units) is equal to the output prediction steps, t_{out}.
- *LSTM:* The model consists of an input layer of 32 long short-term memory (LSTM) units. The output layer consists of a layer similar to the Linear model (described above).
- *GRU:* The model consists of an input layer of 32 gated recurrent units (GRU). The output layer consists of a layer similar to the Linear model (described above).

We initialize the weight metrics of the models as zeros in the beginning. Only the recurrent models (LSTM and GRU) can train and predict based on entire input sequences. The Linear model predicts based on the last input time step. While training, we use the mean squared error (MSE) as the loss function to optimize the model using Adam optimizer [47]. For predictions, we use the mean absolute error (MAE) metric to evaluate model performance. We regularize our training process with the early stopping [43] mechanism, which monitors loss within validation data and stops training if performance does not improve. Based on our tests on different data sets, we configure the training to run for 120 iterations (epochs).

III. METHODS

To describe the methodology, first, we define our spatial units. We designate the term *infection center* (*IC*) to indicate a spatial building block of the model. An *IC* is a point in the space (regional map) where observed or estimated

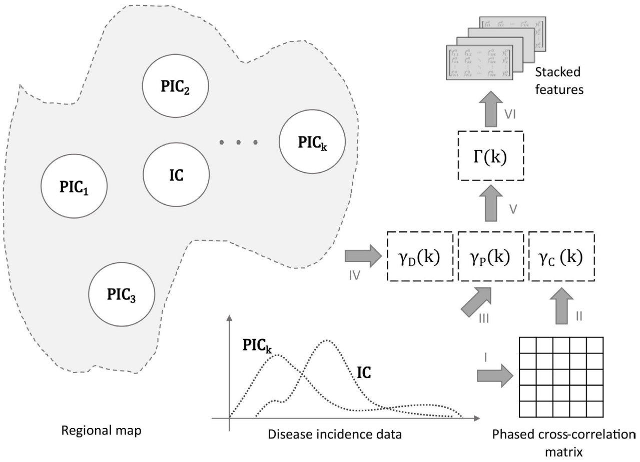

FIGURE 2. A framework depicting the method to expand trainable and testable data set on dengue incidence. The shaded region enclosed by dashed borders on the left depicts a map of the region of interest. The circles inside the map indicate multiple infection centers (*IC*) across the region. The target infection center is marked as *IC*, while the k^{th} peripheral infection center is PIC_k. Using the proposed windowed cross-correlation method (section III-B), a phased cross-correlation matrix is computed (I), which is eventually reduced to a correlation weight, $\gamma_C(k)$ (II). We also compute a prevalence weight, $\gamma_P(k)$, using cumulative case data (III) and a geographic distance weight, $\gamma_D(k)$, using location data (IV). The three weight metrics are combined to compute (V) the predictor metric of PIC_k, $\Gamma(k)$. The *PIC*s are ranked using these Γ values, and their incidence data are selected accordingly to be stacked together with the *IC* feature set (VI).

incidence data on disease outbreaks are available. The spatial granularity of an *IC* is not fixed for the model. It can be a country, a state, a city, a suburb, or an administrative unit with some resolution in space based on available disease incidence data. The basic structure of our proposed framework is shown in Figure 2. The circles inside the shaded region are the infection centers. One of the infection centers, marked as *IC*, indicates the target infection center where we intend to make predictions of a designated label (e.g., dengue cases). The map's remaining infection centers are marked as *peripheral infection centers* (*PIC*). These are locations where similar observations on the designated label of the target *IC* are available. The temporal resolution of the *IC* and the *PIC*s are aligned before performing any comparative analysis of the data. Some locations may have daily, weekly, or monthly observations. Some data may need re-sampling in the time domain before they can be compared (e.g., convert daily weather data to weekly values).

Assuming that there are N peripheral infection centers (*PIC*) on the map. Once we match the spatial and the temporal dimensions of the label data (e.g., weekly dengue cases in a city), we use a windowed-time shifted cross-correlation analysis on each $IC - PIC_k$ pair (for all $k \in N$) and compute a correlation weight, $\gamma_C(k)$. We also consider the cumulative cases of each *PIC* and compute a prevalence weight, $\gamma_P(k)$. Finally, the geodesic distance of each $IC - PIC_k$ pair is taken into account in the form of a distance metric, $\gamma_D(k)$. All three metrics are normalized and lie within the range [0, 1] for the selected region. These are combined as following to compute a predictor strength metric for each PIC_k,

$$\Gamma(k) = \gamma_C(k)[\gamma_P(k) + \gamma_D(k)], \quad \forall k \in N \tag{2}$$

A. WINDOWING INCIDENCE DATA

We compare the time series of disease incidence data (e.g., weekly cases per 100k people) to find correlations. The comparisons are made for all $IC - PIC_k$ pairs with available data for a given region. The key intuition behind this approach is, if a *PIC* in the region has an infectious outbreak at some point in time, $t = t_0$, this may initiate or affect the course of an outbreak for the target *IC* at $t \geq t_0$. A leading *PIC* outbreak may not always imply causal influence depending on the geographic location and population behavior. Despite that, a *PIC* having an outbreak will influence adjacent *IC*s, as it acts as an infection source. This also assumes that a strict isolation measure is not in place and the control measures are not 100% effective due to the vector-borne nature of the infection. It is also important to note that, despite seasonal patterns, outbreaks can occur irregularly. A *PIC* may lead the target *IC* in one season and lag in other seasons due to complex interactions of multiple viral strains. Hence, we divide the time series into smaller time windows. We propose two methods for windowing incidence data: i) fixed-length window allocation and ii) variable-length window detection. Both of these methods are depicted in Figure 3.

A straightforward approach is to divide the entire time series into a fixed number of intervals, M^f. If the time series is T units (day, week, or month) long in total, then fixed window, w_m^f (where $m \in [0, M^f - 1]$) will have T/M^f units of data. While this is the simplest way to divide the series for correlation analysis, it may not be the most efficient. Setting the appropriate value of M_f remains an open problem, although we apply some intuitions from the seasonality patterns of dengue outbreaks in our case. The fixed windows might not be appropriately placed to contain the time series curve's meaningful dynamics, and some windows may even cover regions without an outbreak.

A second approach is to detect windows based on the time series itself. To do this, we normalize the incidence rate of the target *IC* to be ranged between [0, 1]. A typical outbreak curve has irregularities in its shape that make the analysis cumbersome. We use a Savitzky-Golay filter [48] to smooth out the irregularities while preserving the shape of the outbreaks. Our window detection method has two parameters: the incidence threshold (i_{MIN}) and the minimum window size (Δ_{MIN}). An window is detected between two time points, t_{START} and t_{END} (where, $0 \leq t_{START} \leq t_{END} \leq T$), if the normalized incidence rate, $i_N(t) > i_{MIN}$ for all $t_{START} \leq t \leq t_{END}$ and $t_{END} - t_{START} \geq \Delta_{MIN}$. A detected window, w_m^d will have a length (greater than Δ_{MIN}) that depends on the time series curve characteristics. The number of detected windows, M^d, will also vary for the same reason. Figure 3 shows both methods in action using time series data for Brazil's municipality between 2010-2015. For the assigned value of $M^f = 5$, we get equally sized windows, each of

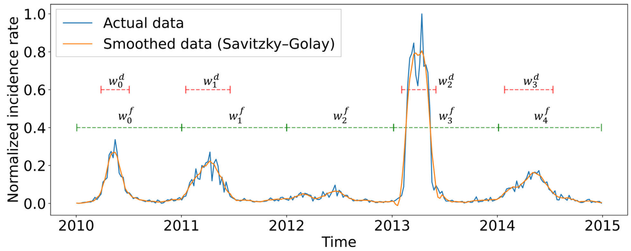

FIGURE 3. Windowing methods used for data segmentation before computing time-shifted correlation coefficients. We present the normalized incidence data from the *Cachoeiro de Itapemirim* municipality of Brazil [42]. The fixed windows (w_m^f) are marked in green, and the detected windows (w_m^d) are marked in red. We set $M^f = 5$ to get 5 fixed windows, each comprising 1 year of data. For the detected windows, we configure $\Delta_{MIN} = 10$ and $i_{MIN} = 0.05$. A Savitzky-Golay [48] smoothing is applied to the data before window detection takes place.

which is 1 year in length. With our detection method, we find 4 windows ($M^d = 4$) that indicate 4 separate outbreaks ($\Delta_{MIN} = 10, i_{MIN} = 0.05$).

Once the windows are selected (either by assignment of fixed number or detection), the following procedures are identical. Hence, we will ignore the superscripts (f and d) in this paper's next sections for brevity. M will depict the total number of windows. w_m (where $m \in [0, M-1]$) will depict the $(m+1)^{th}$ window.

B. TIME-SHIFTED CROSS CORRELATION

Let $i_0(t)$ and $i_k(t)$ be the disease incidence (new cases at time step t) of the target IC and the k^{th} PIC respectively. We perform bivariate computations of time shifted Pearson's correlation coefficients [49] using windowed (w_m) incidence data of the target IC ($i_0^{w_m}(t)$) and the k^{th} PIC ($i_k^{w_m}(t)$). The formula used to compute the coefficients is shown in (3). This measure is also known as time lagged (or phased) cross-correlation [50] in statistics and signal processing.

$$R_k^{w_m}(\theta) = r(i_0^{w_m}(t), i_k^{w_m}(t-\theta)) \tag{3}$$

Here, $R_k^{w_m}(\theta)$ is the correlation coefficient computed for subsets of the time series $i_0(t)$ and $i_k(t)$ selected by the time window w_m when one series is shifted by an amount θ with respect to another. The Pearson's correlation coefficient function is indicated by $r()$ in (3).

For the two time series curves shown in Figure 4, the time-lagged correlation matrix (obtained by computing $R_k^{w_m}(\theta)$ $\forall$ $m \in [0, M-1]$ and $\theta \in [-8, 8]$) is plotted as a heatmap in Figure 5. We use the location depicted in Figure 3 as the target IC ($i_0(t)$) and another location from the same state (Espirito Santo) of Brazil as the k^{th} PIC ($i_k(t)$). For this demonstration, the time series curves were split using fixed length windows ($M^f = M = 5$). The heatmap depicted in Figure 5 can be visually verified by comparing with Figure 4. As expected, the two curves are almost in phase in w_0, negatively correlated in w_2, IC leads in w_1, and PIC leads in w_3 and w_4.

Heatmaps like Figure 5 are computed for all PIC_k with $k \in [1, N]$. To determine if a PIC_k is leading in a window (w_m), we find the location (θ) of the peak correlation as shown in (4).

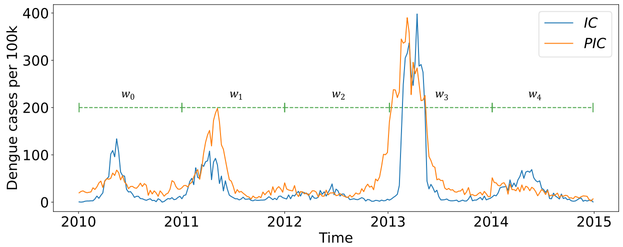

FIGURE 4. Windowing of *IC* and *PIC* incidence data for computing time-shifted cross-correlation coefficients. We allocate a fixed number of windows ($M^f = M = 5$) for the time range 2010-2015, making each window 1 year long (52 weeks approximately). The two curves shown here correspond to the target *IC* ($i_0(t)$) and the k^{th} *PIC* ($i_k(t)$) and these are from the municipalities: *Cachoeiro de Itapemirim* and *Vitória* respectively [42]. The m^{th} time window is marked by the symbol w_m. Note, in the first window (w_0), the outbreaks of *IC* and *PIC* are almost in phase, whereas in the second window (w_1), the *PIC* outbreak is lagging in time compared to the *IC*.

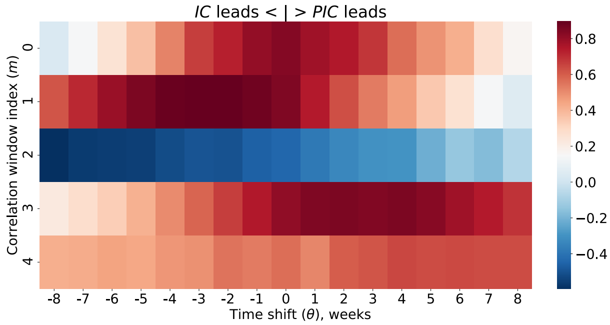

FIGURE 5. Heatmap of the computed time-shifted cross-correlation matrix (3) for the *IC* and the *PIC* in Figure 4. The vertical axis depicts the window indices ($m \in M$) and the horizontal axis depicts time shift (phase), θ that ranges from −8 to +8 weeks. The color shades of the heatmap depict the correlation values demonstrated by the gradient bar on the right. Higher correlation values on the left of the midpoint ($\theta < 0$) indicate that *IC* is leading in the outbreak curve (Figure 4) compared to the *PIC*. Higher correlation values on the right of the midpoint ($\theta > 0$) indicate the opposite (*PIC* leads *IC*).

$$\theta_k^{w_m} = \underset{\theta}{\text{argmax}}\, R_k^{w_m}(\theta) \tag{4}$$

We define the correlation strength $S_k^{w_m}$ to be the mean correlation measure computed around the peak ($\theta_k^{w_m}$), extending by the amount θ_E in both directions (5).

$$S_k^{w_m} = \frac{1}{\sum_{\theta=-\theta_E}^{\theta_E} 1} \sum_{\theta=-\theta_E}^{\theta_E} R_k^{w_m}(\theta_k^{w_m} + \theta) \tag{5}$$

Equation (5) has an additional condition on the values θ such that $R_k^{w_m}(\theta_k^{w_m} + \theta)$ exists for the given parameters. A PIC_k leads the IC if the peak of correlation lies on the right half of the heatmap shown in Figure 5, which translates to $\theta_k^{w_m} > 0$. We only consider if a PIC_k is at least in phase with the IC and discard the cases where any PIC_k lags the IC. This is how we compute the predictor probability matrix P with dimensions $M \times N$. The individual predictor probabilities ($\forall m \in [0, M-1]$, $\forall k \in [1, N]$) are computed as shown in (6).

$$P_{m,k} = \begin{cases} S_k^{w_m}, & \text{if } \theta_k^{w_m} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The predictor probabilities are averaged across all windows (7) to compute overall predictive abilities of all PIC_k. For a particular region ($k \in [1, N]$), the predictive ability metrics are normalized between [0,1]. The final measure, $\gamma_C(k)$, is defined as the *correlation weight* of PIC_k as shown in (8).

$$\hat{\gamma_C}(k) = \frac{1}{M} \sum_{m=0}^{M-1} P_{m,k} \tag{7}$$

$$\gamma_C(k) = \frac{\hat{\gamma_C}(k) - \min_k \hat{\gamma_C}(k)}{\max_k \hat{\gamma_C}(k) - \min_k \hat{\gamma_C}(k)} \tag{8}$$

C. DISTANCE AND PREVALENCE METRICS

With increasing distance, the impact of a PIC on the target IC is likely to be reduced due to decreased travel between the locations, increasing differences in environmental conditions (e.g., temperature, rainfall, vegetation), etc. It is intuitive to use a metric proportional to the inverse distance for strengthening the predictive ability measures of PICs. Let, d_k be the geodesic distance [51] (shortest path on the surface of the earth, assuming earth to be an ellipsoid) between the target IC and PIC_k. The normalized [0, 1] distance of a PIC_k in the region is,

$$\hat{d_k} = \frac{d_k - \min_k d_k}{\max_k d_k - \min_k d_k} \tag{9}$$

We want the metric to be inversely proportional to the distance. Hence, the distance metric of PIC_k is defined as,

$$\gamma_D(k) = 1 - \hat{d_k} \tag{10}$$

The outbreak history of a location is an important criterion that indicates the viral pathogen and endemic scenarios' persistence. For a PIC_k, we compute the prevalence I_k by taking a sum of the incidence data $i_k(t)$ for the entire timeline ($\forall t \in [0, T]$).

$$I_k = \sum_{t=0}^{T} i_k(t) \tag{11}$$

The prevalence metric is normalized [0, 1] across the region.

$$\gamma_P(k) = \frac{I_k - \min_k I_k}{\max_k I_k - \min_k I_k} \tag{12}$$

IV. RESULTS

We present here the results in several stages. The outcomes of the feature analysis are presented first. This is followed by an analysis of prediction performance using the proposed methods. The results are generated using municipality-wise weekly dengue case data between 2010-2019 from Brazil's Espírito Santo state. We obtained data for 78 municipalities of Espírito Santo and ranked them based on the total number of cases recorded for the entire time period. The top 5 municipalities based on prevalence are listed in Table 2. The location IDs shown in the table are the IBGE (Instituto Brasileiro de Geografia e Estatística) codes for Brazil [52].

TABLE 2. Top 5 locations of Espírito Santo ranked by total reported cases of dengue during 2010-2019 [42].

Loc. ID	Name	Cases	Cases per 100k
3205309	Vitória	71,348	19,501.72
3205002	Serra	58,424	11,081.10
3201209	Cachoeiro de Itapemirim	45,319	21,520.12
3205200	Vila Velha	36,743	7,329.18
3201308	Cariacica	27,103	7,059.60

TABLE 3. Top 5 predictor *PIC*s for Vitória. The weights based on our defined predictability metrics (correlation, prevalence, and distance) are shown in columns 3-5. The combined weights (Γ) are shown in the last column.

Rank #	Loc. ID	Corr. γ_C	Prev. γ_P	Dist. γ_D	Γ
1	3201209	0.912	0.812	0.614	1.301
2	3205200	1.000	0.270	1.000	1.270
3	3201308	0.956	0.260	0.996	1.202
4	3205101	0.621	0.770	0.937	1.060
5	3200607	0.772	0.524	0.809	1.029

A. FEATURE SELECTION AND ANALYSIS

The PICs are sorted and ranked for each IC, based on the predictor strength metric, Γ, as shown in (2). This is computed from the three individual metrics: correlation weight (γ_C), prevalence weight (γ_P), and distance weight (γ_D). We choose the municipality of Vitória in Espirito Santo, Brazil, as the target IC, which had the highest total number of cases in the state during 2010-2019, to generate the results. We compute the correlation weight using 20 fixed-length windows ($M^f = M = 20$) for the time range 2010-2019, making each window approximately 26 weeks (6 months) long. The top 5 ranked PICs based on Γ are listed in Table 3 with the corresponding weights. While our method prioritizes the correlation weight more than others, PICs with relatively lower correlation weight can still be favored because of the following factors: i) having a significant number of cases or ii) being in proximity of the target IC. This is evident in Table 3 as 3201209 (PIC #1) is chosen over 3205200 (PIC #2) and 3205101 (PIC #4) is chosen over 3200607 (PIC #5). *Note, the numbers (#) used in 'PIC #' indicate rank. This should not be confused with arbitrary indices (k) used to compute metrics (PIC_k).* This effect is also illustrated in Figure 6, which shows the ranked PICs listed in Table 3. Although PIC #1 lies farthest from the IC among the five (lowest γ_D), it is ranked at the top due to significantly higher values in the other two factors (γ_C and γ_P).

A time series plot in Figure 7 shows that the top PICs are mostly correlated with the IC, Vitória. Among the PICs displayed here, PIC #4 (3205101) shows the weakest correlation with the IC. It will be clear in the upcoming results, the proximity and the high incidence fraction of this location help with prediction performance. The variability of the incidence curves prevents our method from classifying a single

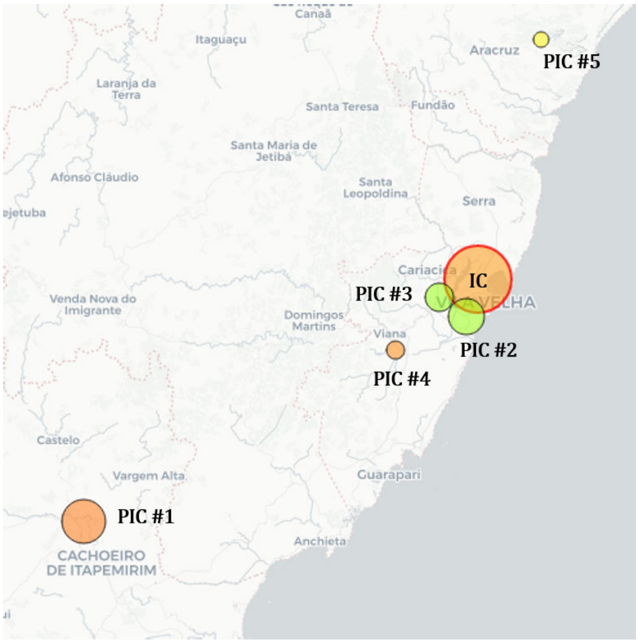

FIGURE 6. A geospatial map showing an incidence center (*IC*) and 5 ranked peripheral incidence centers (*PIC*). The circles' radii are proportional to the total number of cases reported during 2010-2019 [42]. The shades of the fill color are generated from a color gradient (green-yellow-red) which is proportional to the fraction of cases with respect to the local population of each location (*IC* or *PIC*) during 2010-2019 [42]. The greenish shades indicate smaller infected fractions, while the reddish shades indicate larger fractions. Map generated using Folium [53] with OpenStreetMap [54]. Basemap tiles provided by CartoDB [55].

PIC as the optimal predictor for all time ranges. However, the combination of top-ranked *PIC*s will improve prediction accuracy.

B. PREDICTION PERFORMANCE

After selecting features with the proposed methods, we train and evaluate the prediction performances using the three models (Linear, LSTM, and GRU) described in section II-E. The time series data are split into 3 distinct sets with ratios of 50:30:20 for model training, evaluation, and testing. A sliding window with input length (t_{in}) of 8 and output length (t_{out}) 4 is used for a single shot prediction of the next 4 weeks using data of the past 8 weeks every step.

1) INDIVIDUAL *IC* PREDICTION

For the *IC* of Vitória, the *PIC*s are added gradually according to their computed ranks (section IV-A), and the models' prediction performances are evaluated. The mean absolute error (MAE) values on the normalized test data are plotted in Figure 8 for varying number of additional features, N_{PIC}. For both LSTM and GRU models, the addition of the first two PICs deteriorates the model performance. However, the subsequent additions keep improving the outcomes. The plots quickly reach their minima, after which errors increase. The first few additions increase error due to high variability in the incidence data, as evident in Figure 7. Further additions of *PIC* create averaging effects on the predictor data set and cause performance improvements. The GRU model eventually reaches a minimum MAE value of 0.128, which is lower than the best optimal LSTM prediction (0.1415) by about 9.54%. The Linear model reaches an optimum MAE value of 0.3384, which is nowhere close to the recurrent models. After reaching the minima, all three error curves rise again. This increase in MAE with larger feature sets (N_{PIC}) can be attributed to the curse of dimensionality [56], [57]. LSTM and GRU models perform optimally with 4 and 6 additional *PIC*s, respectively. Predicting based on present input and historical context (internal states of LSTM and GRU) of data certainly puts recurrent models ahead in performance, which is evident even without a *PIC* ($N_{PIC} = 0$ in Figure 8) in the feature set. However, the linear model significantly benefits from feature addition as case data from *PIC*s strengthen inductive bias.

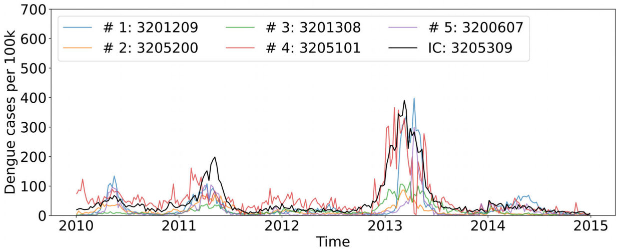

FIGURE 7. Time series plots of weekly dengue cases per 100,000 people for the *IC* and top 5 ranked *PIC*s (Table 3). The plots depict cases only between 2010-2015 for improved clarity, but the metrics were computed based on the entire series (2010-2019).

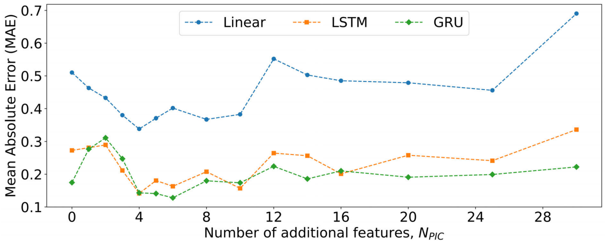

FIGURE 8. Prediction performances of the Linear, LSTM, and GRU models in predicting normalized test data for varying number of features (N_{PIC}). *PIC* data are added to the feature set based on ranks dictated by computed predictor strengths (Table 3), after which models are trained and evaluated over the test data to compute the mean absolute error (MAE) metrics (lower is better).

After determining the optimum number of features (N_{PIC}) to be added to the predictor data set for an *IC*, recurrent models are trained with the extended feature set and are used to predict dengue cases for the test data subset of the time series. Figure 9 compares the predictions with actual incidence data for both LSTM and GRU models using Vitória as the *IC* and choosing the optimum number of *PIC*s for the two models ($N_{PIC} = 4$ for LSTM and 6 for GRU).

2) EFFECT OF WINDOW SELECTION METHODS

The impacts of various correlation window configurations are analyzed under the two proposed windowing methods: i) fixed-length window assignment and ii) variable-length

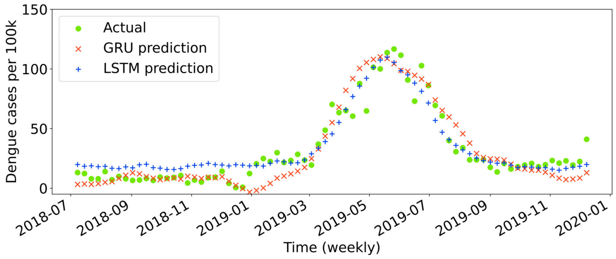

FIGURE 9. **Predicted weekly dengue cases per 100,000 people using test data with the recurrent models (LSTM and GRU). These results were obtained for the *IC*, Vitória, with 4 and 6 additional *PIC* data joined with LSTM and GRU's feature set, respectively.**

window detection. For the first method (fixed), we vary the number of windows (M^f) between 5, 10, 20, and 40. For the second method (detection), we vary the minimum window size (Δ_{MIN}) between 5, 10, 20, and 30 weeks. In both methods, we are consequently varying the number of windows, window lengths, and where the windows are located. In total, we run tests under 8 different scenarios and compare the prediction performances using the 3 models (Linear, LSTM, and GRU). Instead of working with a single IC, we run the tests over multiple locations and report the average performance metrics (e.g., mean of MAE). We take the top 20 ICs based on total cases per 100,000 people and average the performance metrics across ICs. Among the 20 ICs, there were 4 ICs where none of the models could predict with reasonable accuracy (MAE $<$ 1) with or without additional features. In those locations, either the data were too limited or the outbreaks were too random for our models to generalize beyond training data. We filter out these 4 locations and take the remaining 16 locations to evaluate our methods.

For a model, predicting on a given IC, the optimal MAE is defined as the minimum mean absolute error (MAE) obtained by varying the number of additional features (N_{PIC}) from the PIC ranked list produced by a given windowing method (i.e., minima of the curves in Figure 8 are the optimal MAEs of three models). The average optimal MAE (across 16 ICs) are plotted in Figure 10 for all 8 windowing schemes. The recurrent models perform significantly better than the Linear model: on average, LSTM and GRU outperform the Linear model by 60% and 61%, respectively. LSTM and GRU outperform the best-performing linear model by 41% and 43%, respectively. We expect this kind of advantage, given the importance of considering historical data in predicting future dengue cases. The performances across different windowing schemes with LSTM and GRU models are comparable (both models have an approximate standard deviation of 0.02 around their means of 0.3569 and 0.3435, respectively), with a small trend of increasing mean error values with windowing method configurations. In the case of linear models, there are some stark contrasts when using window detection methods. Using the window detection method with a large Δ_{MIN} value increases performance sharply. For example, using a Δ_{MIN} of 30 weeks shows 41.9% improvement over fixed window schemes' average performance.

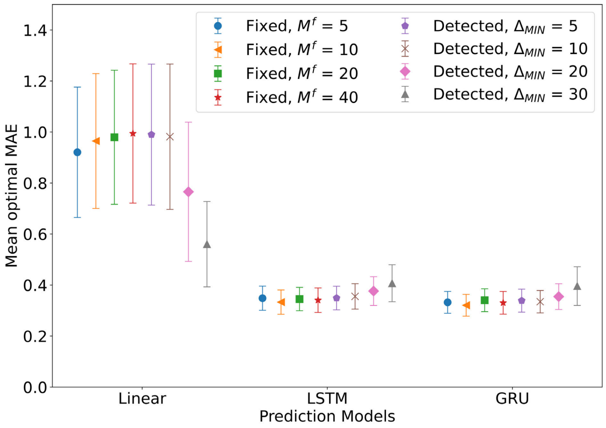

FIGURE 10. **A comparison chart of the lowest prediction error obtainable using different fixed and detected windowing schemes. Four fixed windowing schemes with the number of windows, M^f = 5, 10, 20, and 40, along with four window detection schemes having minimum window lengths, Δ_{MIN} = 5, 10, 20, and 30 weeks are compared for three models. The vertical axis denotes the average of the optimal MAE values. The markers denote the mean values, with the bars representing standard errors of the mean. The results are the averages of 16 top *IC*s based on prevalence.**

This indicates that features selected with a small number of large correlation windows detected based on outbreak locations in the time series are more effective than the remaining schemes that use a relatively large number of smaller windows. A similar trend is also visible with fixed window methods, although it is not statistically conclusive. The lower values of M^f, which translates to having a lower number of large windows, show slightly better performance over higher values of M^f.

To understand how beneficial our proposed methods are over the baseline (without feature enhancement, $N_{PIC} = 0$), we analyze the improvements in prediction accuracy. The term, *improvement in accuracy* is defined as the relative decrease in MAE (i.e., the increase in prediction performance) with optimum PIC selection (Figure 10) compared to MAE without additional features ($N_{PIC} = 0$). We compute the accuracy improvements across all 16 ICs and plot the *mean improvement in accuracy* as percentages in Figure 11. The linear model demonstrates average improvements ranging from 18.97% to 27.13% depending on the window selection schemes. The LSTM model improvements range from 22.13% to 33.6% and the GRU model range from 10.79% to 31.92% depending on the window selection schemes. We should be careful in interpreting these values; greater improvements do not translate to better optimal performance. These values are relative to their baselines ($N_{PIC} = 0$). With some incidence centers (IC), a model can already predict with higher accuracy than other incidence centers. Our methods help minimize those gaps and still provide some improvements over the baselines (ranging from 10.79% to 33.6% depending on the model and the scheme). Figure 10 demonstrates that, in general, GRU and LSTM both perform well when we deal with average values. For individual ICs, however, a conclusive determination of the best performing model can be done (either GRU or LSTM).

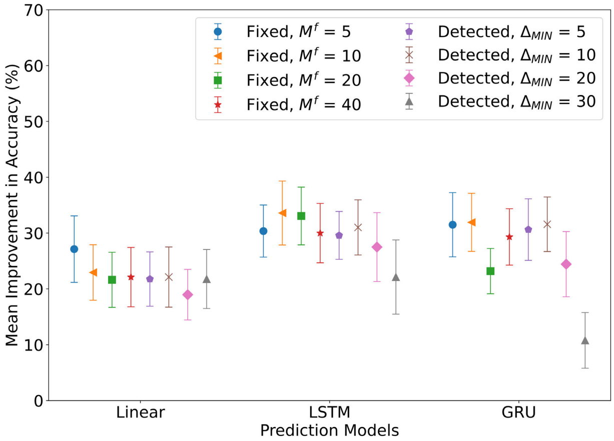

FIGURE 11. A comparison chart of average improvement in accuracy (reduction in MAE) with respect to performance without feature enhancements ($N_{PIC} = 0$) for different fixed and detected windowing schemes. Four fixed windowing schemes with the number of windows, $M^f = 5$, 10, 20, and 40, along with four window detection schemes having minimum window lengths, $\Delta_{MIN} = 5$, 10, 20, and 30 weeks are compared for three models. The markers denote the mean values, with the bars representing standard errors of the mean. The results are the averages of 16 top *IC*s based on prevalence.

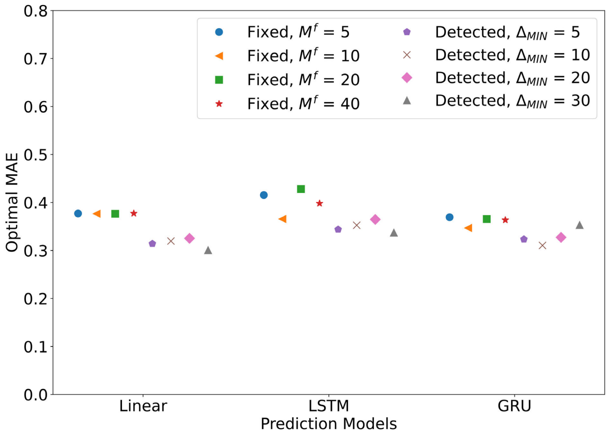

FIGURE 12. A comparison chart of lowest prediction error obtainable using different fixed and detected windowing schemes for case data aggregated across an ecoregion. Four fixed windowing schemes with the number of windows, $M^f = 5$, 10, 20, and 40, along with four window detection schemes having minimum window lengths, $\Delta_{MIN} = 5$, 10, 20, and 30 weeks are compared for three models. The vertical axis denotes the optimum mean absolute error (MAE). The results were obtained for the ecoregion named Bahia coastal forest [59].

3) PERFORMANCE ON AGGREGATED DATA

It is sometimes reasonable to aggregate data to larger scales based on geographic adjacency and environmental similarities (e.g., weather). From a macroscopic point of view, predicting for an ecoregion may be more meaningful for policymakers to interpret the outcomes. According to Omernik (2004), ecoregions are defined as areas within which there is a spatial coincidence in characteristics of geographical phenomena (e.g., geology, physiography, vegetation, land use, climate, hydrology, terrestrial and aquatic fauna, etc.) associated with differences in the quality, health, and integrity of ecosystems [58]. We use the terrestrial ecoregions defined by The Nature Conservancy(TNC) in this work [59]. The following analysis is performed for Brazilian locations with available data in the Bahia Coastal Forest ecoregion.

Comparing the prediction performance on aggregated data shows that the advantages of recurrent models compared to the Linear model (which we had for individual *IC*s) are diminished. The optimal performances across prediction models become more comparable, as shown in Figure 12. The mean optimal MAE values obtained for 4 fixed window schemes are 0.38, 0.40, and 0.36 when using Linear, LSTM, and GRU models, respectively. The mean optimal MAEs for 4 window detection schemes are 0.31, 0.35, and 0.33 using the same three models. We observe a distinct advantage of the window detection method for selecting *PIC*s. On average, the window detection methods improve over their baselines ($N_{PIC} = 0$) by 16.50%, 12.68%, and 10.07% for Linear, LSTM, and GRU models, respectively. Variable window allocation based on outbreak window detection outperforms fixed window allocation methods. In other words, windowed cross-correlation on a few important outbreak regions performs better than comparing over the entire time series. As the target data is aggregated from all *PIC*s in the region, the optimally trained Linear models sometimes perform better than recurrent models. One key takeaway is that cases for the entire region can be predicted with improved accuracy while using a

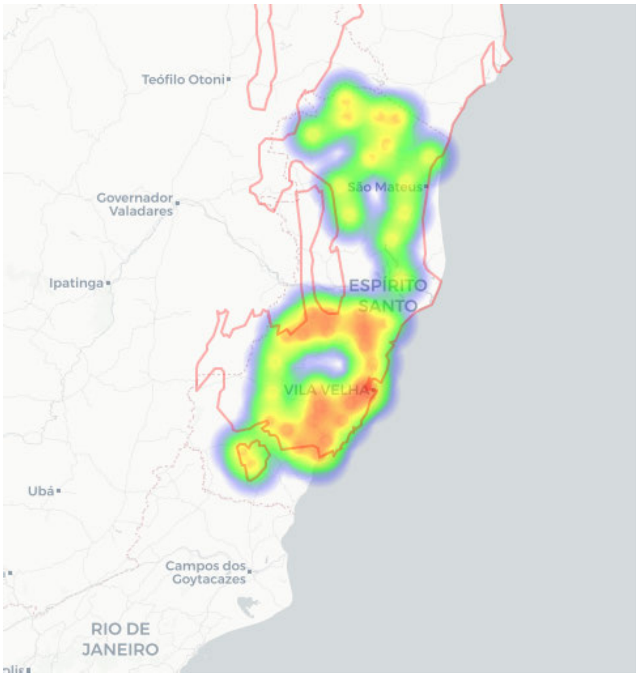

FIGURE 13. A geospatial map showing the predicted risk of infection in an ecoregion of Brazil. A part of the ecoregion (Bahia Coastal Forests) shown here is marked with red borders. The risk of infection is shown as a heatmap with color shades ranging from blue (low risk) to red (high risk). The heatmap is constructed by combining the ranked *PIC*s in the ecoregion, with higher-ranked *PIC*s contributing more to the risk. The aggregated data was predicted using the Linear model with an optimum number of features ($N_{PIC} = 4$) selected with window detection method ($\Delta_{MIN} = 30$). This heatmap depicts the risk on the date of 09-June-2019, with 40 top-ranked *PIC*. Map generated using Folium [53] with OpenStreetMap [54]. Basemap tiles provided by CartoDB [55].

small subset (PICs) of data as features. For the results shown in Figure 12, the optimal MAE values can be reached for N_{PIC} values between 2 and 5. This aggregation provides a fast way to predict cases, with a small fraction of regional outbreak data. While the accuracy achieved at the ecoregion level is not on par with the accuracy achieved for single ICs, this prediction is useful to generate risk maps.

We construct a risk map for the Bahia Coastal Forest ecoregion using the predicted cases and weights of the top-ranked PIC in Figure 13. The top 40 PICs are included in the map, and the PICs that contribute more towards the weight (Γ) are shown as high-risk regions (e.g., red). The aggregated prediction was done using the optimal Linear model with $\Delta_{MIN} = 30$ and $N_{PIC} = 4$.

V. CONCLUSION

In this work, we develop a method to select relevant incidence data from peripheral locations as features to improve the prediction of dengue fever outbreaks. In order to rank features, we use windowed cross-correlation analysis on dengue case data. We propose two methods for allocating correlation windows (position and size) over the time series to compute correlation weights. For a target location (IC), peripheral locations (PIC) are ranked based on a combination of correlation, distance, prevalence metrics. The predictive models benefit from the ranked feature sets by achieving model and location-specific optimal performances with a relatively small subset of features. We tested three predictive models using dengue case data from Brazil, showing different levels of accuracy gains.

On average, the proposed feature enhancement methods improve prediction performance by 10.79% to 33.6% over the baseline feature set for the locations we tested, depending on the prediction model and the window allocation scheme. For the location with the highest total cases (2010-2019) in the Espírito Santo region of Brazil, we could get MAE values as low as 0.13 (normalized case data) using the GRU model with data from just 6 locations added to the feature set. In a test across multiple locations, both RNN models (LSTM and GRU) performed with comparable accuracy (average MAE ranging from 0.3435 to 0.3569) when using an optimal number of additional features. The Linear model also benefited (18.97% to 27.13% improvement over the baseline) from windowed correlation-based feature enhancements, although its performance never got close to recurrent models. Compared using the respective optimal number of features, the best performing recurrent models outperformed the Linear model by at least 41% in prediction accuracy. The window detection methods showed performance comparable to fixed window allocation. This similarity is advantageous when working with extensive data sets, as the detected windows use a smaller subset of data than fixed allocation. For municipality-level dengue case prediction, GRU was the best performing model, closely followed by LSTM. The performance gaps between these two models diminished after feature set optimization. When predicting aggregated data for the entire region using a subset of constituent locations, our methods reach optimal performance with the addition of only 2-5 locations (out of 77), depending on the model and window selection scheme. This is especially useful for situations where the lack of trainable data hinders forecasting. For example, dengue risk for a region can be predicted if a small subset of data from epidemically important region locations is available. Aggregated prediction addresses incomplete data and eases the 'curse of dimensionality' while improving training efficiency.

Future efforts in this area can focus on gaining further insights based on location's epidemiological, environmental, and economic characteristics and combining them to improve feature ranks. While case data are indicators of the severity of an outbreak, these are not always adequate to explain future possibilities. Machine learning models cannot generalize beyond training data for every location with the same degree of accuracy. Complex interactions of dengue viral strains and changes in host immunity patterns may evolve outbreak characteristics over the years. Several random factors may affect dengue outbreaks, including host travel patterns, natural disasters, and lifestyle changes because of other infectious outbreaks (e.g., COVID-19 pandemic). While the metrics used in our method capture such factors' long-term characteristics, these do not account for randomness. While keeping the feature sets reasonably small, our method can improve outbreak prediction. The proposed method can be generalized and used for projecting any infectious outbreak where temporal data at a reasonable spatial resolution are available.

CODE AVAILABILITY

Software libraries were developed in Python for the framework presented in this paper. The libraries, along with example code and data, can be found on GitHub using the link: *https://github.com/tanvir-ferdousi/dengue-machine-learning*.

DISCLAIMER

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The conclusions in this report are those of the authors and do not necessarily represent the views of the USDA. USDA is an equal opportunity provider and employer.

REFERENCES

[1] R. E. Baker, J.-M. Peña, J. Jayamohan, and A. Jérusalem, "Mechanistic models versus machine learning, a fight worth fighting for the biological community?" *Biol. Lett.*, vol. 14, no. 5, May 2018, Art. no. 20170660.

[2] 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, and M. F. Myers, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[3] G. A. Roth, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, and A. J. T. L. Abdelalim, "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017," *Lancet*, vol. 392, no. 10159, pp. 1736–1788, 2018.

[4] R. Li, L. Xu, O. N. Bjørnstad, K. Liu, T. Song, A. Chen, B. Xu, Q. Liu, and N. C. Stenseth, "Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 9, pp. 3624–3629, Feb. 2019.

[5] D. O. Fuller, A. Troyo, and J. C. Beier, "El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica," *Environ. Res. Lett.*, vol. 4, no. 1, Jan. 2009, Art. no. 014011.

[6] P.-C. Wu, J.-G. Lay, H.-R. Guo, C.-Y. Lin, S.-C. Lung, and H.-J. Su, "Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan," *Sci. Total Environ.*, vol. 407, no. 7, pp. 2224–2233, Mar. 2009.

[7] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee, "Distribution, seasonal variation & dengue transmission prediction in Sisaket, Thailand," *Indian J. Med. Res.*, vol. 138, no. 3, p. 347, 2013.

[8] M. P. Mammen, C. Pimgate, C. J. M. Koenraadt, A. L. Rothman, J. Aldstadt, A. Nisalak, R. G. Jarman, J. W. Jones, A. Srikiatkhachorn, C. A. Ypil-Butac, A. Getis, S. Thammapalo, A. C. Morrison, D. H. Libraty, S. Green, and T. W. Scott, "Spatial and temporal clustering of dengue virus transmission in Thai villages," *PLoS Med.*, vol. 5, no. 11, p. e205, Nov. 2008.

[9] A. González-Vidal, V. Moreno-Cano, F. Terroso-Sáenz, and A. F. Skarmeta, "Towards energy efficiency smart buildings models based on intelligent data analytics," *Proc. Comput. Sci.*, vol. 83, pp. 994–999, Jan. 2016.

[10] M. V. Moreno, F. Terroso-Saenz, A. Gonzalez-Vidal, M. Valdes-Vela, A. F. Skarmeta, M. A. Zamora, and V. Chang, "Applicability of big data techniques to smart cities deployments," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 800–809, Apr. 2017.

[11] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, "Using causal discovery for feature selection in multivariate numerical time series," *Mach. Learn.*, vol. 101, nos. 1–3, pp. 377–395, Oct. 2015.

[12] Y. Hmamouche, A. Casali, and L. Lakhal, "A causality based feature selection approach for multivariate time series forecasting," in *Proc. 9th Int. Conf. Adv. Databases, Knowl., Data Appl. (DBKDA)*, 2017, pp. 1–7.

[13] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 141–144, Jun. 2003.

[14] N. Jin, Y. Zeng, K. Yan, and Z. Ji, "Multivariate air quality forecasting with nested long short term memory neural network," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8514–8522, Dec. 2021.

[15] S. F. Crone and N. Kourentzes, "Feature selection for time series prediction—A combined filter and wrapper approach for neural networks," *Neurocomputing*, vol. 73, nos. 10–12, pp. 1923–1936, Jun. 2010.

[16] L. Munkhdalai, T. Munkhdalai, K. H. Park, T. Amarbayasgalan, E. Erdenebaatar, H. W. Park, and K. H. Ryu, "An end-to-end adaptive input selection with dynamic weights for forecasting multivariate time series," *IEEE Access*, vol. 7, pp. 99099–99114, 2019.

[17] D. Liu, L. Clemente, C. Poirier, X. Ding, M. Chinazzi, J. Davis, A. Vespignani, and M. Santillana, "Real-time forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models," *J. Med. Internet Res.*, vol. 22, no. 8, Aug. 2020, Art. no. e20285.

[18] E. Mussumeci and F. Codeço Coelho, "Large-scale multivariate forecasting models for Dengue–LSTM versus random forest regression," *Spatial Spatio-Temporal Epidemiol.*, vol. 35, Nov. 2020, Art. no. 100372.

[19] D. Schoenherr, J. Paulick, B. M. Strauss, A.-K. Deisenhofer, B. Schwartz, J. A. Rubel, W. Lutz, U. Stangier, and U. Altmann, "Identification of movement synchrony: Validation of windowed cross-lagged correlation and -regression with peak-picking algorithm," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0211494.

[20] M. Hu, X. Feng, Z. Ji, K. Yan, and S. Zhou, "A novel computational approach for discord search with local recurrence rates in multivariate time series," *Inf. Sci.*, vol. 477, pp. 220–233, Mar. 2019.

[21] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, vol. 734. Hoboken, NJ, USA: Wiley, 2011.

[22] A. E. Permanasari, I. Hidayah, and I. A. Bustoni, "SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of malaria incidence," in *Proc. Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Oct. 2013, pp. 203–207.

[23] W.-C. Hong, "Application of seasonal SVR with chaotic immune algorithm in traffic flow forecasting," *Neural Comput. Appl.*, vol. 21, no. 3, pp. 583–593, 2012.

[24] T. W. Scott, E. Chow, D. Strickman, P. Kittayapong, R. A. Wirtz, L. H. Lorenz, and J. D. Edman, "Blood-feeding patterns of *Aedes aegypti* (Diptera: Culicidae) collected in a rural Thai village," *J. Med. Entomol.*, vol. 30, no. 5, pp. 922–927, Sep. 1993.

[25] S. Shrivastava, D. Tiraki, A. Diwan, S. K. Lalwani, M. Modak, A. C. Mishra, and V. A. Arankalle, "Co-circulation of all the four dengue virus serotypes and detection of a novel clade of DENV-4 (genotype I) virus in Pune, India during 2016 season," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192672.

[26] J. H. Huber, M. L. Childs, J. M. Caldwell, and E. A. Mordecai, "Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission," *PLOS Neglected Tropical Diseases*, vol. 12, no. 5, May 2018, Art. no. e0006451.

[27] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee, "Impact of human mobility on the emergence of dengue epidemics in Pakistan," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 38, pp. 11887–11892, Sep. 2015.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at Uber," in *Proc. ICML*, vol. 34, 2017, pp. 1–5.

[30] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep learning with long short-term memory for time series prediction," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 114–119, Jun. 2019.

[31] L. Zhu and N. Laptev, "Deep and confident prediction for time series at Uber," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 103–110.

[32] J. Gu, L. Liang, H. Song, Y. Kong, R. Ma, Y. Hou, J. Zhao, J. Liu, N. He, and Y. Zhang, "A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[33] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212.

[34] J. Xu, K. Xu, Z. Li, F. Meng, T. Tu, L. Xu, and Q. Liu, "Forecast of dengue cases in 20 Chinese cities based on the deep learning method," *Int. J. Environ. Res. Public Health*, vol. 17, no. 2, p. 453, Jan. 2020.

[35] S. R. Christophers, *Aedes Aegypti: Yellow Fever Mosquito*. Cambridge, U.K.: CUP Archive, 1960.

[36] J. Couret, E. Dotson, and M. Q. Benedict, "Temperature, larval diet, and density effects on development rate and survival of *Aedes aegypti* (Diptera: Culicidae)," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e87468.

[37] H. Delatte, G. Gimonneau, A. Triboire, and D. Fontenille, "Influence of temperature on immature development, survival, longevity, fecundity, and gonotrophic cycles of *Aedes albopictus*, vector of chikungunya and dengue in the Indian Ocean," *J. Med. Entomol.*, vol. 46, no. 1, pp. 33–41, Jan. 2009.

[38] W. A. Rowley and C. L. Graham, "The effect of temperature and relative humidity on the flight performance of female *Aedes aegypti*," *J. Insect Physiol.*, vol. 14, no. 9, pp. 1251–1257, 1968.

[39] L. B. Carrington, S. N. Seifert, N. H. Willits, L. Lambrechts, and T. W. Scott, "Large diurnal temperature fluctuations negatively influence *Aedes aegypti* (Diptera: Culicidae) life-history traits," *J. Med. Entomol.*, vol. 50, no. 1, pp. 43–51, Jan. 2013.

[40] D. M. Watts, R. E. Whitmire, D. S. Burke, A. Nisalak, and B. A. Harrison, "Effect of temperature on the vector efficiency of *Aedes aegypti* for dengue 2 virus," *Amer. J. Tropical Med. Hygiene*, vol. 36, no. 1, pp. 143–152, Jan. 1987.

[41] M. A. Johansson, F. Dominici, and G. E. Glass, "Local and global effects of climate on dengue transmission in Puerto rico," *PLoS Neglected Tropical Diseases*, vol. 3, no. 2, p. e382, Feb. 2009.

[42] I. Dengue. (2021). *Info Dengue Surveillance Project*. [Online]. Available: https://info.dengue.mat.br/

[43] L. Prechelt, "Early stopping—But when?" in *Neural Networks: Tricks Trade* (Lecture Notes in Computer Science), vol. 7700, G. Montavon, G. B. Orr, and K. R. Müller, Eds. Berlin, Germany: Springer, 2012, pp. 55–69, doi: 10.1007/978-3-642-35289-8_5.

[44] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.

[45] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[46] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[49] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing* (Springer Topics in Signal Processing), vol. 2. Berlin, Germany: Springer, 2009, doi: 10.1007/978-3-642-00296-0_5.

[50] (2021). *Cross Correlation*. [Online]. Available: https://en.wikipedia.org/wiki/Cross-correlation

[51] C. F. F. Karney, "Algorithms for geodesics," *J. Geodesy*, vol. 87, no. 1, pp. 43–55, Jan. 2013.

[52] (2021). *Instituto Brasileiro de Geografia e Estatística*. [Online]. Available: https://www.ibge.gov.br/

[53] Python-Visualization. (2021). *Folium*. [Online]. Available: https://python-visualization.github.io/folium/

[54] O. Foundation. (2021). *OpenStreetMap*. [Online]. Available: https://www.openstreetmap.org

[55] C. Inc. (2021). *CartoDB Positron*. [Online]. Available: https://carto.com/

[56] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[57] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Jan. 2018.

[58] J. M. Omernik, "Perspectives on the nature and definition of ecological regions," *Environ. Manage.*, vol. 34, no. S1, pp. S27–S38, Apr. 2004.

[59] D. M. Olson and E. Dinerstein, "The global 200: Priority ecoregions for global conservation," *Ann. Missouri Botanical Garden*, vol. 89, pp. 199–224, Apr. 2002.

TANVIR FERDOUSI (Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, in 2013, and the Ph.D. degree in electrical and computer engineering from Kansas State University, Manhattan, KS, USA, in 2021. He was a Senior Software Engineer at Samsung Research and Development Institute Bangladesh (SRBD), Dhaka, from 2013 to 2016. He is currently a Postdoctoral Research Associate with the University of Virginia Biocomplexity Institute. His research interests include network theory, machine learning, and predictive models.

LEE W. COHNSTAEDT received the Ph.D. degree in epidemiology and public health from Yale University, in 2008. He is currently a Research Entomologist with the Agricultural Research Service for the United States Department of Agriculture, where he works on vector-borne diseases of medical and agricultural importance. His current interests include disease vector biting midges, and mosquitoes and insect population management methods to reduce pathogen transmission.

CATERINA M. SCOGLIO (Senior Member, IEEE) received the Dr.Eng. degree from the Sapienza University of Rome, Italy, in 1987. She is currently a Paslay Chair Professor of electrical and computer engineering with Kansas State University, where her main research interests include the field of network science and engineering. Before joining Kansas State University, she worked at the Fondazione Ugo Bordoni, from 1987 to 2000, and the Georgia Institute of Technology, from 2000 to 2005. She is also affiliated as a Faculty Member with the Institute of Computational Comparative Medicine (ICCM), Kansas State University. Her main research interest includes the modeling and analysis of complex networks, with applications in epidemic spreading and power grids.