AWARD NUMBER: DWFP grant W911QY-19-1-0004

TITLE: PICTUREE: Predicting Insect Contact and Transmission Using histoRical Entomological and Environmental data

PRINCIPAL INVESTIGATOR: Caterina Scoglio

CONTRACTING ORGANIZATION: Kansas State University
Manhattan, KS

REPORT DATE: 30 September 2023

TYPE OF REPORT: Final

PREPARED FOR: Sponsoring Agency

DISTRIBUTION STATEMENT: Approved for Public Release
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
20230930	Final	20190701	20230630

4. TITLE AND SUBTITLE

PICTUREE: Predicting Insect Contact and Transmission Using histoRical Entomological and Environmental data

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
W911QY191004		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
0011371761		

6. AUTHOR(S)

Caterina Scoglio, Aram Vajdi, Chunlin Yi, Lee Cohnstaedt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Kansas State University
Fairchild Hall Manhattan, KS

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
W6QK ACC-APG NATICK BLDG1 General Greene Avenue, Natick, MA		

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Dengue fever remains a significant public health concern in many tropical and subtropical countries, and there is still a need for a system that can effectively combine global risk assessment with timely incidence forecasting. This research describes an integrated application called PICTUREE—Aedes, which can collect and analyze dengue-related data, display simulation results, and forecast outbreak incidence. PICTUREE—Aedes automatically updates global temperature and precipitation data and contains historical records of dengue incidence (1960–2012) and Aedes mosquito occurrences (1960–2014) in its database. The application utilizes a mosquito population model to estimate mosquito abundance, dengue reproduction number, and dengue risk. To predict future dengue outbreak incidence, PICTUREE—Aedes applies various forecasting techniques, including the ensemble Kalman filter, recurrent neural network, particle filter, and super ensemble forecast, which are all based on user-entered case data. The PICTUREE—Aedes' risk estimation identifies favorable conditions for potential dengue outbreaks, and its forecasting accuracy is validated by available outbreak data from Cambodia.

15. SUBJECT TERMS

Dengue, prediction, estimation, forecast, risk, mosquito

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	C. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
U	U	U	UU	90

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Caterina Scoglio	7853173154

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE.
Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE.
State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED.
Indicate the time during which the work was performed and the report was written.

4. TITLE.
Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER.
Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER.
Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER.
Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER.
Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER.
Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER.
Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/ FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

Introduction 3
The PICTUREE project 3
Development of a global spatiotemporal dengue risk 3
Kalman filter forecast 3
Machine learning forecast 4
The PICTUREE web application 4
Key research accomplishments 4
Reportable outcomes 5
Conclusion 6
References 6
Appendices 7

Introduction

Vector-borne diseases result from, for example, viruses, parasites, or bacteria transmitted to humans from the bite of an infected mosquito, biting fly, sand fly, or tick arthropod vector [1], [2]. The US military repeatedly suffers mission degradation from vector-borne disease during contingency operations, with medically-related reductions in personnel ranging from 10-50% depending on the theater of operation and available countermeasures [3]. The global distribution of joint US military forces supports faster and larger deployments to numerous diverse environments, complicating exposure risks. The Deployed War-Fighter Protection research program has invested in tools and products that protect troops from nuisance and disease-vector arthropods and increase combat readiness [4]. A decision support tool was needed to optimize location-specific provisioning, alignment, prophylaxis/treatment, personal protection, vector control, and disease/vector surveillance resources based on the estimated risk for transmission of vector-borne pathogens. Here, we have developed an innovative and unique system to close this capability gap, the PICTUREE (Predicting Insect Contact and Transmission Using histoRical Entomological and Environmental data) logistical support tool, focusing on *Aedes aegypti* and dengue [5]. Estimating current and future dengue transmission risk will reduce waste of limited shelf-life products and increase logistical efficiency to the hub and forward areas. Planners can use PICTUREE software to determine a Service Member's likely current and future risk based on geographic location, the deployment season, and estimated time in the country. Furthermore, the software allows commanders to monitor in-country conditions to decide when and which products to (re)order, and, following redeployment, medical providers can evaluate likely past exposure to help with clinical diagnosis.

The PICTUREE project

In the following, we describe the most relevant components of our work for the entire research period, adding the complete publications related to those achievements in the appendices.

Development of a global spatiotemporal dengue risk [A, B]

Given that the life cycle of Aedes mosquitoes is strongly influenced by habitat temperature, numerous studies have utilized temperature-dependent entomological parameters of these mosquitoes to construct virus transmission and outbreak risk models [6], [7]. In this study, we advance existing research by developing a mechanistic model for the mosquito life cycle that accurately accounts for the non-Markovian nature of the process. Furthermore, we demonstrate how to reduce the model to the corresponding differential equations, enabling us to utilize existing methods for analyzing the system and fitting the model to observations. This approach can be further applied to similar non-Markovian processes. By fitting the model to data on human dengue cases, we estimate several model parameters, allowing the development of a global spatiotemporal dengue risk map. This risk model employs temperature and precipitation data to assess the environmental suitability for dengue outbreaks in any given area.

Kalman filter forecast [C, D]

We developed a new model for dengue outbreak estimation and forecast that adopts the vector-borne disease model SEIR-SEI with compartments Susceptible-Exposed-Infectious-Recovered (for host) and Susceptible-Exposed-Infectious (for vector) into the ensemble Kalman filtering (EnKF) assimilation method [8]. The SEIR-SIR-EnKF model was first validated using synthetic dengue outbreak in twin experiments. Then, the model produced good performance when applied to estimate and forecast the dengue outbreak dynamics with real historical time-series cases in 3 different cities. Furthermore, we

compared the accuracy of the real-time predictions between SEIR-SEI-EnKF model, SEIR-EnKF model, and SIR-EnKF model; we found the SEIR-SEI-EnKF model had the most accurate predictions.

The ensemble Kalman filter was also used to forecast COVID-19 cases in the Hubei province, and it performed well as a short-term predictor of the daily cases [9].

Machine learning forecast [E]

The performance of data-driven models depends on training samples. For accurately predicting dengue fever cases, historical incidence data are inadequate in many locations [10]. This work enhances temporally limited dengue case data by the methodological addition of epidemically relevant case data from nearby places as predictors (features). A novel framework is presented for windowing incidence data and computing time-shifted correlation-based metrics to quantify feature relevance [11]. 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.

The PICTUREE web application [F]

This research has produced an integrated application called PICTUREE—Aedes, which can collect and analyze dengue-related data, display simulation results, and forecast outbreak incidence[5]. PICTUREE—Aedes automatically updates global temperature and precipitation data and contains historical records of dengue incidence (1960–2012) and Aedes mosquito occurrences (1960–2014) in its database [12], [13]. The application utilizes a mosquito population model to estimate mosquito abundance, dengue reproduction number, and dengue risk. To predict future dengue outbreak incidence, PICTUREE—Aedes applies various forecasting techniques, including the ensemble Kalman filter, recurrent neural network, particle filter, and super ensemble forecast, which are all based on user-entered case data. The PICTUREE—Aedes' risk estimation identifies favorable conditions for potential dengue outbreaks, and its forecasting accuracy is validated by available outbreak data from Cambodia.

Public access to the PICTUREE—Aedes application is provided at the following URL: http://picturee.engg.ksu.edu.

Key research accomplishments

- Non-Markovian mathematical models of *Aedes aegypti* population and dengue transmission
- Accurate dengue risk map
- Ensemble Kalman filter predictor
- Neural network predictor
- Particle filter predictor
- Super ensemble predictor
- The PICTUREE-Aedes web application

- The Nepal government has used the PICTUREE application to forecast dengue cases in regions of their country.

Reportable outcomes

Published and submitted papers:

A. Aram Vajdi, Lee W. Cohnstaedt, and Caterina M. Scoglio. "Assessing Dengue Risk Globally Using Non-Markovian Models." Submitted for journal publication (2023)
B. Riad, Mahbubul H., Lee W. Cohnstaedt, and Caterina M. Scoglio. "Risk Assessment of dengue transmission in Bangladesh using a spatiotemporal network model and climate data." *The American journal of tropical medicine and hygiene* 104, no. 4 (2021): 1444
C. Yi, Chunlin, Lee W. Cohnstaedt, and Caterina M. Scoglio. "SEIR-SEI-EnKF: A new model for estimating and forecasting dengue outbreak dynamics." *IEEE Access* 9 (2021): 156758-156767.
D. Yang, Qihui, Chunlin Yi, Aram Vajdi, Lee W. Cohnstaedt, Hongyu Wu, Xiaolong Guo, and Caterina M. Scoglio. "Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China." *Infectious Disease Modelling* 5 (2020): 563-574.
E. Ferdousi, Tanvir, Lee W. Cohnstaedt, and Caterina M. Scoglio. "A Windowed Correlation-Based Feature Selection Method to Improve Time Series Prediction of Dengue Fever Cases." *IEEE Access* 9 (2021): 141210-141222.
F. Yi, Chunlin, Aram Vajdi, Tanvir Ferdousi, Lee W. Cohnstaedt, and Caterina Scoglio. "PICTUREE—Aedes: A Web Application for Dengue Data Visualization and Case Prediction." *Pathogens* 12, no. 6 (2023): 771.

Developed software and databases (stored on GitHub and available on request):

- Software:
  - PICTUREE-Aedes web application,
  - ensemble Kalman filter predictor,
  - neural network predictor,
  - particle filter predictor,
  - super ensemble predictor.
- Databases:
  - weather station data,
  - occurrence data (dengue and vectors),
  - daily weather observation (average temperature, maximum temperature, minimum temperature, and precipitation),
  - estimated dengue reproduction number,
  - estimated *Aedes* mosquito population,
  - estimated dengue risk.

Video on YouTube where Chunlin Yi, Ph.D. student, explains how to use the PICTUREE-Aedas web application, available at https://www.youtube.com/watch?v=Sufl7Lh5ibs.

PhD degrees obtained and postdocs supported:

- PhD students who graduated and were supported by this project: Aram Vajdi, Md Mahbubul Huq Riad, Tanvir Ferdousi, and Chunlin Yi
- Postdoc: Dr. Aram Vajdi

Conclusion

We have proposed the PICTUREE logistical support tool to get a snapshot of current and future risks from *dengue* in any world region. This dynamic platform can help the US military to be operationally prepared for this vector-borne disease based on geographic location and environmental conditions. The platform can also be helpful for medical providers to assess the past risk profile of a previously deployed service member for targeted post-operation medical treatments. This unique platform processes a region's relevant historical and current data to forecast current and future mosquito and dengue pathogen exposure risks. Optimal risk forecasts are estimated using computational models through an array of filters, such as ensemble Kalman filters, particle filters, deep neural networks, and their super-ensemble that combine forecasts into a collective prediction that is more robust than individual forecast models. PICTUREE provides a dynamic and flexible platform capable of adapting to all possible locations.

This prototype of PICTUREE has a user-friendly interface to show the current and future risks of dengue using easy-to-understand maps accessible through a web application at http://picturee.engg.ksu.edu. Other programs or apps created by the DWFP to help planners with vector-related issues can be integrated with PICTUREE to obtain a single comprehensive tool. PICTUREE has focused on the mosquito *Aedes aegypti* and its associated dengue pathogen as a proof of concept and to create and validate a core system. Future work includes extending the core PICTUREE tool to additional vectors/pathogen systems, including *Culex* and West Nile virus. Overall, our PICTUREE tool summarizes data into a simplified visual image that supports a proactive and adaptive approach to preventing pathogen transmission rather than reacting to outbreaks and case reports.

References

[1] "Vector-borne diseases." https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases (accessed Sep. 21, 2023).

[2] S. C. Weaver and W. K. Reisen, "Present and future arboviral threats," *Antiviral research*, vol. 85, no. 2, pp. 328–345, 2010.

[3] K. E. Jones *et al.*, "Global trends in emerging infectious diseases," *nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[4] L. Eisen and C. G. Moore, "*Aedes* (Stegomyia) *aegypti* in the continental United States: a vector at the cool margin of its geographic range," *Journal of medical entomology*, vol. 50, no. 3, pp. 467–478, 2013.

[5] C. Yi, A. Vajdi, T. Ferdousi, L. W. Cohnstaedt, and C. Scoglio, "PICTUREE—Aedes: A Web Application for Dengue Data Visualization and Case Prediction," *Pathogens*, vol. 12, no. 6, p. 771, 2023.

[6] L. Lambrechts *et al.*, "Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7460–7465, 2011.

[7] M. H. Riad, L. W. Cohnstaedt, and C. M. Scoglio, "Risk Assessment of dengue transmission in Bangladesh using a spatiotemporal network model and climate data," *The American journal of tropical medicine and hygiene*, vol. 104, no. 4, p. 1444, 2021.

[8] C. Yi, L. W. Cohnstaedt, and C. M. Scoglio, "SEIR-SEI-EnKF: A new model for estimating and forecasting dengue outbreak dynamics," *IEEE Access*, vol. 9, pp. 156758–156767, 2021.

[9] Q. Yang *et al.*, "Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China," *Infectious Disease Modelling*, vol. 5, pp. 563–574, 2020.

[10] A. Wesolowski *et al.*, "Impact of human mobility on the emergence of dengue epidemics in Pakistan," *Proceedings of the national academy of sciences*, vol. 112, no. 38, pp. 11887–11892, 2015.
[11] T. Ferdousi, L. W. Cohnstaedt, and C. M. Scoglio, "A Windowed Correlation-Based Feature Selection Method to Improve Time Series Prediction of Dengue Fever Cases," *IEEE Access*, vol. 9, pp. 141210–141222, 2021.
[12] M. U. Kraemer *et al.*, "The global compendium of *Aedes aegypti* and *Ae. albopictus* occurrence," *Scientific data*, vol. 2, no. 1, pp. 1–8, 2015.
[13] J. P. Messina, O. J. Brady, D. M. Pigott, J. S. Brownstein, A. G. Hoen, and S. I. Hay, "A global compendium of human dengue virus occurrence," *Scientific Data*, vol. 1, no. 1, pp. 1–6, 2014.

Appendices

Assessing Dengue Risk Globally Using Non-Markovian Models

Aram Vajdi* , Lee W. Cohnstaedt[1,‡] and Caterina M. Scoglio*

Abstract

Dengue is a vector-borne disease transmitted by Aedes mosquitoes. The worldwide spread of these mosquitoes and the increasing disease burden have emphasized the necessity for a spatio-temporal risk map capable of assessing the environmental suitability for dengue outbreaks. Given that the life cycle of Aedes mosquitoes is strongly influenced by habitat temperature, numerous studies have utilized temperature-dependent entomological parameters of these mosquitoes to construct virus transmission and outbreak risk models. In this study, we advance existing research by developing a mechanistic model for the mosquito life cycle that accurately accounts for the non-Markovian nature of the process. Furthermore, we demonstrate how to reduce the model to the corresponding differential equations, enabling us to utilize existing methods for analyzing the system and fitting the model to observations. This approach can be further applied to similar non-Markovian processes. By fitting the model to data on human dengue cases, we estimate several model parameters, allowing the development of a global spatiotemporal dengue risk map. This risk model employs temperature and precipitation data to assess the environmental suitability for dengue outbreaks in a given area.

Index Terms

Non-Markovian Models, Dengue, Aedes mosquitoes

1 Introduction

Dengue fever, a vector-borne viral disease transmitted primarily by Aedes mosquitoes, presents a significant global public health challenge [1]. The main carrier of the dengue virus is the Aedes aegypti mosquito, an urban species whose life cycle is strongly influenced by environmental temperature. Additionally, Aedes aegypti serves as the primary vector for the Chikungunya virus and the Zika virus, while Aedes albopictus plays a secondary role [2], [3], [4]. The global prevalence of Aedes mosquitoes, along with its expansion into new regions [5], necessitates the development of outbreak risk models which are crucial for guiding the control strategies aimed at eliminating virus transmission.

In recent years, there has been a growing interest in developing mechanistic models that integrate entomological and environmental components to gain a better understanding of the spread of dengue [6], [7], [8], [9], [10], [2], [11], [12], [13], [14], [15]. For instance, Yang et al. conducted a study in which they measured entomological parameters of Aedes aegypti, including temperature-dependent development and mortality rates [7]. They incorporated these parameters into differential equations that describe dengue transmission within a transmission model. Using this mechanistic model, they determined that a temperature of $29^{\circ}\mathrm{C}$ is the most favorable for a dengue outbreak [8], [9]. In another study [2], the authors utilized a calibrated basic reproductive number for dengue to develop an outbreak risk model. In the field of mathematical epidemiology, the reproductive number is defined as the number of secondary infected cases generated by one infected case in a healthy population. Therefore, if this number is greater than one, the number of new cases increases. Nevertheless, it's crucial to highlight that even though both of these studies incorporate temperature-dependent biological parameters, their mathematical equations and subsequent derivations rest upon the assumption of a constant environmental temperature. In another noteworthy study [16], researchers compiled a global database of Aedes mosquito occurrences and employed pertinent environmental variables—such as temperature susceptibility, precipitation, and urbanization—as predictive factors to develop a niche model that defines the global suitability map for Aedes mosquito presence.

In this paper, we present a novel mechanistic model that captures the intricate relationships between vector populations and environmental variables, allowing for the development of a more accurate global dengue risk map. In the majority of existing mechanistic models for mosquito-borne diseases, the mosquito life cycle is divided into multiple phases, including aquatic stages and the adult stage. Subsequently, by leveraging experimental data on developmental and mortality rates at each of these stages, a set of differential equations is formulated to provide a mathematical description of the mosquito population dynamics. For example, using parameters α and β to represent the development rates of mosquito larvae and pupae, and μ to denote the mortality rate of pupae, the following equation:

$$\frac{dp}{dt} = \alpha l - (\beta + \mu)p$$

• [1] *Corresponding author*
[*] *Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS*
[‡] *United States Department of Agriculture, Agricultural Research Service, Foreign Arthropod-Borne Animal Diseases Research Unit, Manhattan, Kansas*

has been employed to estimate the mosquito pupae population represented by variable p, with the larvae population being denoted as l. However, even in an environment with a constant temperature and consequently a constant rate, such an equation does not precisely describe the mosquito life cycle. This discrepancy arises from the fact that this equation describes Markovian transitions with exponential sojourn times, while in reality, this is not always the case. In fact, a new pupa emerges when the larval development is complete, and this development is not necessarily exponentially distributed; it often occurs around a specific time that depends on the temperature. Therefore, the actual process is non-Markovian. Interestingly, these considerations have been taken into account in a previous study on Aedes mosquito population dynamics, where simulations were employed to accommodate non-Markovian transitions [17].

While most traditional epidemiological compartment models assume exponentially distributed sojourn times, there exist studies that have considered relaxing this assumption to offer a more precise representation of transmission dynamics [18], [19], [20]. One effective approach is to approximate the sojourn time distribution with an Erlang distribution [21], [18]. This is because a random variable with an Erlang distribution can be seen as the sum of multiple exponentially distributed random variables. Consequently, by partitioning an epidemiological state into several phases characterized by exponential transition times, traditional compartment models can provide an approximate description of the non-Markovian process. Indeed, rather than relying solely on Erlang distributions, an even more accurate approximation can be achieved by utilizing a broader class of distributions known as phase-type distributions [19].

In this paper, we present a novel mechanistic model aimed at describing dengue transmission. Our model leverages the temperature-dependent development rates of mosquitoes more accurately compared to existing models. Drawing inspiration from a recent study [18] on non-Markovian epidemiological models, we initially formulate integral equations for the non-Markovian dynamics of the mosquito life cycle. Subsequently, we illustrate how assuming a phase-type distribution for transitions simplifies the integral equations into ordinary differential equations (ODEs). While this model is specifically designed for dengue transmission, our approach can be extended to other relevant epidemiological systems.

One of the primary objectives of this paper is to develop a global, time-dependent risk metric that quantifies the environmental suitability for dengue outbreaks. Here, we utilize available epidemic data from various locations around the world to create this risk map. Interestingly, our map designates a high risk to areas that are not identified as high risk in other dengue risk maps. To accomplish this, we fit the outbreak data to the dengue transmission model using a particle filter algorithm, enabling the estimation of the environment's carrying capacity and mosquito biting rates. Subsequently, we develop a numerical risk model that incorporates these estimated values, along with daily average temperature and precipitation, to calculate the global spatiotemporal dengue outbreak risk based on the dengue reproductive number and vector population.

The structure of the remaining sections of this paper is as follows: Section 2 presents the integral equations used to describe the mosquito life cycle, emphasizing the non-Markovian nature of transitions between different life stages. Subsequently, we demonstrate the reduction of these integral equations to corresponding ordinary differential equations (ODEs). Moving forward, we incorporate dengue virus transmission into the mosquito life cycle model, resulting in the virus transmission model. In Section 3, we outline the methodology employed for developing an environment carrying capacity model based on precipitation and available outbreak data. Furthermore, we detail the construction of our risk model, which is based on the transmission model and the estimated parameters. Lastly, Section 4 concludes the paper by summarizing the contributions of our novel mechanistic model and discussing potential avenues for future research

2 MOSQUITOES LIFE CYCLE MODEL

To determine the population of mosquitoes at various stages of growth, we use a model with four states representing the egg, larva, pupa, and adult stages. Figure 1 illustrates the different states and the corresponding rates in the model. It is well established that the development time for each life-cycle stage is temperature dependent and the transition to the next stage occurs after the development is completed [17], [22], [23], [7], [9], [14], [24], [2]. Therefore, calculating the mosquito population requires tracking the extent of development across the life-cycle stages. We adopt the following master equations to determine the total population in each state of the life cycle.

$$
\begin{aligned}
l(t) &= \int_0^t \lambda_{e,l}(u)\ \mathcal{P}_{l,p}(u,t)\ \mathcal{P}_{l,d}(u,t)\ du + l(0)\ \mathcal{P}_{l,p}(0,t)\ \mathcal{P}_{l,d}(0,t) \\
p(t) &= \int_0^t \lambda_{l,p}(u)\ \mathcal{P}_{p,a}(u,t)\ \mathcal{P}_{p,d}(u,t)\ du + p(0)\ \mathcal{P}_{p,a}(0,t)\ \mathcal{P}_{p,d}(0,t) \\
a(t) &= \int_0^t \lambda_{a}(u)\ \mathcal{P}_{a,e}(u,t)\ \mathcal{P}_{a,d}(u,t)\ du + a(0)\ \mathcal{P}_{a,e}(0,t)\ \mathcal{P}_{a,d}(0,t) \\
e(t) &= \int_0^t \lambda_{e}(u)\ \mathcal{P}_{e,l}(u,t)\ \mathcal{P}_{e,d}(u,t)\ du + e(0)\ \mathcal{P}_{e,l}(0,t)\ \mathcal{P}_{e,d}(0,t)
\end{aligned}
\tag{1}
$$

Here, we define $l(t)$, $p(t)$, $a(t)$, and $e(t)$ as the quantities representing the number of larvae, pupae, adult female mosquitoes, and eggs, respectively, irrespective of their developmental stage. The term $\lambda_{e,l}(u)$ represents the rate of larval emergence at time u. In other words, $\lambda_{e,l}(u)du$ corresponds to the number of eggs that hatch into larvae during the time interval $[u, u+du]$. Furthermore, $\mathcal{P}_{l,p}(u,t)$ denotes the probability that a larva, which emerged at time u, remains in the

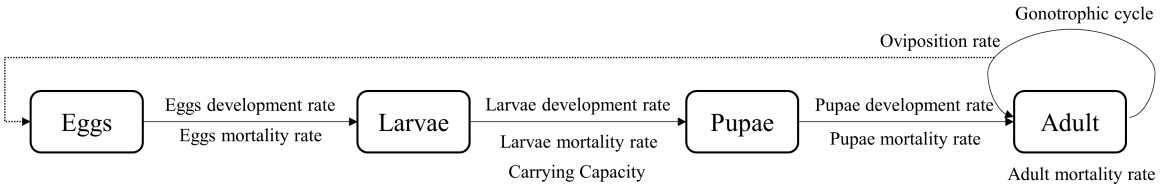


Fig. 1. The Aedes mosquitoes life cycle. The rates depicted in the figure are temperature-dependent, which consequently change over time.

larval stage and has not yet transformed into a pupa by time t, given that it has survived up to time t. This probability is unity unless the larva completes its development, at which point it becomes zero. Assuming the larval development rate at temperature T is denoted as $\gamma_{l,p}(T)$, we can express the extent of development from time u to t as

$$\mathcal{D}_{l,p}(u,t) = \int_u^t \gamma_{l,p}(T(\tau))\, d\tau, \tag{2}$$

where $\mathcal{D}_{l,p} = 1$ indicates full development. This equation aligns with an approximation used to calculate the duration of the development period when the environmental temperature is variable [reference]. With the development extent quantified as we have described, we can formulate the probability of the larval stage continuing from time u to time t as

$$\mathcal{P}_{l,p}(u,t) = \mathcal{S}_{l,p}(\mathcal{D}_{l,p}(u,t)), \tag{3}$$

where $\mathcal{S}_{l,p}$ represents the survival function of a probability density function $f(x)$ centered around one, i.e.,

$$\mathcal{S}_{l,p}(\mathcal{D}) = 1 - \int_0^D f(x)\, dx.$$

Figure 2 shows the survival function for two different choices of density functions. The black curve is survival function for the gamma distribution with shape parameter 100 and rate 0.01. Considering this survival function as $\mathcal{S}_{l,p}(\mathcal{D})$, implies the probability of transitioning to the pupa stage when the larvae development is less than 0.9 is smaller than 0.14, and this probability increases to 0.84 when the development reaches 1.1. This is in agreement with the simulation method in [17]. In contrast, utilizing the survival function of the exponential distribution with a rate of 1, depicted by the red curve in the figure, results in 59 percent of larvae progressing to the pupal stage before their development reaches 0.9. In fact, we will demonstrate that describing the mosquitoes' life cycle using traditional ODEs with temperature-dependent rates is equivalent of employing the survival function of exponential distribution in the equations 1. Thus, for a more accurate representation of the life cycle, we utilize the integral equations 1 with survival functions similar to the black curve shown in Figure 2. Similar to the development rate, the mortality rate of larvae, $\gamma_{l,d}$, is temperature dependent. Consequently, we express the survival probability of the larvae, $\mathcal{P}_{l,d}(u,t)$, as

$$\mathcal{P}_{l,d}(u,t) = \mathcal{S}_{l,d}(\mathcal{A}_{l,p}(u,t)),\ \mathcal{A}_{l,d}(u,t) = \int_u^t \gamma_{l,d}(T(\tau))\, d\tau. \tag{4}$$

To maintain the consistency in the presentation of equations 1, we write the larvae survival probability in terms of a general survival function $\mathcal{S}_{l,d}$, but in our calculations, we use the exponential distribution survival function, which is $\mathcal{S}(x) = e^{-x}$. This choice of exponential distribution implies the assumption that the mortality of larvae is a non-homogeneous markovian process with a temperature dependent mortality rate.

The master equations 1 include the emergence rates as well as survival probabilities for the pupa, adult, and egg stages, which are defined in a manner similar to the probabilities and rates for the larva stage. To complete the life cycle description we need to express the emergence rates of each stage in terms of the parameters of the preceding stage. To achieve this, we analyze the derivative of $l(t)$ with respect to time to determine the pupal emergence rate. It is straight forward to show

$$\begin{aligned}\frac{dl(t)}{dt} =& \lambda_{e,l}(t)\ \mathcal{P}_{l,p}(t,t)\ \mathcal{P}_{l,d}(t,t) \\ &+ \int_0^t \lambda_{e,l}(u)\ \frac{\partial \mathcal{P}_{l,p}(u,t)}{\partial t}\ \mathcal{P}_{l,d}(u,t)\ du - l(0)\ \frac{\partial \mathcal{P}_{l,p}(0,t)}{\partial t}\ \mathcal{P}_{l,d}(0,t) \\ &+ \int_0^t \lambda_{e,l}(u)\ \mathcal{P}_{l,p}(u,t)\ \frac{\partial \mathcal{P}_{l,d}(u,t)}{\partial t}\ du - l(0)\ \mathcal{P}_{l,p}(0,t)\ \frac{\partial \mathcal{P}_{l,d}(0,t)}{\partial t}.\end{aligned} \tag{5}$$

The equation above demonstrates that the larvae rate of change consists of three components: the first line on the right-hand side of the equation represents the influx from the egg stage, while the remaining two lines represent the outflow to the pupae and dead larvae stages, respectively. Taking into account that $\mathcal{P}_{l,p}(t,t) = 1$ and $\mathcal{P}_{l,d}(t,t) = 1$, the inflow at time t is $\lambda_{e,l}(t)$, which is consistent with our initial definition of $\lambda_{e,l}$. Additionally, $-dt\partial\mathcal{P}_{l,p}(u,t)/\partial t$ is the probability that larvae that emerged at time u transition to the pupal stage within the time interval $[t, t+dt]$, given their survival until time t.

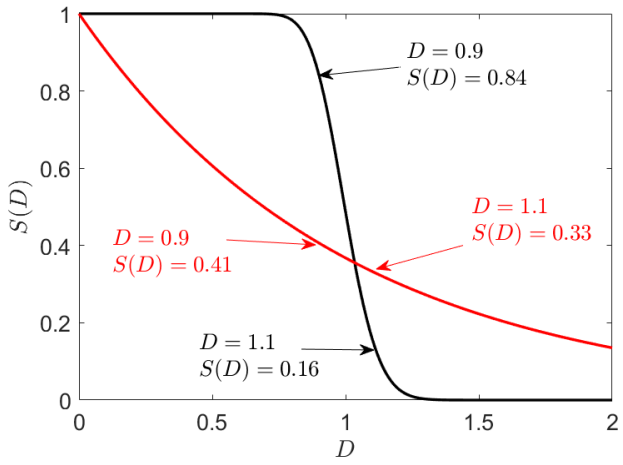


Fig. 2. The survival function of the Erlang distribution is represented by the black curve, while that of the exponential distribution is depicted by the red curve.. Both distributions have a mean equal to one.

Similarly, $-\partial \mathcal{P}_{l,d}(u,t)/\partial t$ represents the mortality probability density of the larvae at time t. Indeed, if we assume $\mathcal{P}_{l,d}$ is written in terms of the survival function for the exponential distribution we obtain $-\partial \mathcal{P}_{l,d}(u,t)/\partial t = \mathcal{P}_{l,d}(u,t)\gamma_{l,d}(T(t))$.

Thus, our analysis of the terms in the expression for $dl(t)/dt$ demonstrates that the second line in equation 5 represents the emergence rate of pupae at time t. With a similar reasoning, the emergence rates in the equations 1 can be expressed as follows

$$
\begin{aligned}
\lambda_{l,p}(\tau) &= -\int_0^\tau \lambda_{e,l}(u)\ \frac{\partial \mathcal{P}_{l,p}(u,\tau)}{\partial \tau}\ \mathcal{P}_{l,d}(u,\tau)\ du - l(0)\ \frac{\partial \mathcal{P}_{l,p}(0,\tau)}{\partial \tau}\ \mathcal{P}_{l,d}(0,\tau) \\
\lambda_a(\tau) &= \frac{1}{2}\lambda_{p,a}(\tau) + \lambda_{a,a}(\tau) \\
\lambda_{p,a}(\tau) &= -\int_0^\tau \lambda_{l,p}(u)\ \frac{\partial \mathcal{P}_{p,a}(u,\tau)}{\partial \tau}\ \mathcal{P}_{p,d}(u,\tau)\ du - p(0)\ \frac{\partial \mathcal{P}_{p,a}(0,\tau)}{\partial \tau}\ \mathcal{P}_{p,d}(0,\tau) \\
\lambda_{a,a}(\tau) &= -\int_0^\tau \lambda_a(\tau)\ \frac{\partial \mathcal{P}_{a,e}(u,\tau)}{\partial \tau}\ \mathcal{P}_{a,d}(u,\tau)\ du - a(0)\ \frac{\partial \mathcal{P}_{a,e}(0,\tau)}{\partial \tau}\ \mathcal{P}_{a,d}(0,\tau) \\
\lambda_e(\tau) &= ov(T(\tau))\lambda_{a,a}(\tau) \\
\overline{\lambda}_{e,l}(\tau) &= -\int_0^\tau \lambda_e(u)\ \frac{\partial \mathcal{P}_{e,l}(u,\tau)}{\partial \tau}\ \mathcal{P}_{e,d}(u,\tau)\ du - e(0)\ \frac{\partial \mathcal{P}_{e,l}(0,\tau)}{\partial \tau}\ \mathcal{P}_{e,d}(0,\tau) \\
\lambda_{e,l}(\tau) &= \left(1 - \frac{l(\tau)}{C(\tau)}\right)\overline{\lambda}_{e,l}(\tau)
\end{aligned}
\tag{6}
$$

We can see that the adult mosquitoes emergence rate, λ_a, is the sum of $1/2\lambda_{p,a}$, which is the inflow from pupa state to adult female state, and $\lambda_{a,a}$ that represents the number adult female mosquitoes completed a gonotrophic cycle and oviposition and starting a new gonotrophic cycle. Here, we assume that half of the pupae population are female, and the other half are male. This two terms are illustrated in figure 1. Additionally, we assume that the number of eggs laid within a time interval $[\tau, \tau + d\tau]$, denoted by $d\tau\lambda_e(\tau)$, is the product of two factors: the number of adult mosquitoes that have completed the gonotrophic cycle within that interval, $d\tau\lambda_{a,a}(\tau)$, and the average number of eggs laid per adult female mosquito at the temperature $T(\tau)$, denoted as $ov(T(\tau))$. We also account for the environment's carrying capacity, denoted as $C(\tau)$, by constraining the number of eggs that can hatch into the larval state. This constraint is reflected in the expression for $\lambda_{e,l}$.

2.1 Describing the mosquitoes life cycle using ODEs

Here we show how to reduce the life cycle integral equations 1 to ODEs when the probabilities $\mathcal{P}(u,t)$ in these equations are given in terms of the survival function of phase-type distributions. A phase-type distribution is the probability distribution for the absorption time in a continuous-time Markov chain with one absorbing state[25], [19]. It is known that any positive-valued distribution can be approximated using a phase-type distribution. For instance, a random variable X that follows an Erlang distribution can be exactly represented as a phase-type distribution. In this case, X can be perceived as the time taken to reach an absorbing state when a continuous-time Markov process with a number of successive transient states and one final absorbing state is initiated from the first state. Formally, a phase-type distribution with m transient states is defined by a pair $(\boldsymbol{\alpha}, \boldsymbol{\mathcal{Q}})$. The $1 \times m$ row vector $\boldsymbol{\alpha}$ is the initial distribution of the Markov chain across the m transient

states and $(\boldsymbol{\mathcal{Q}})$ is an $m \times m$ matrix of transition rates among these transient states. Therefore, the generator matrix $\boldsymbol{G}$ for the Markov chain is given by

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{\mathcal{Q}}_{m\times m} & \boldsymbol{q}_{m\times 1} \\ \boldsymbol{0}_{1\times m} & 0 \end{pmatrix},\ \boldsymbol{q} = -\boldsymbol{\mathcal{Q}}\ \boldsymbol{1},$$

where $\boldsymbol{1}$ is a column vector with all entries set to 1. The cumulative distribution, $\mathcal{C}$, and survival function, $\mathcal{S}$, for a phase-type distribution are

$$\mathcal{C}(t) = 1 - \boldsymbol{\alpha}\ e^{t\boldsymbol{\mathcal{Q}}}\ \boldsymbol{1},\ \mathcal{S}(t) = \boldsymbol{\alpha}\ e^{t\boldsymbol{\mathcal{Q}}}\ \boldsymbol{1}$$

To clarify the procedure of converting the integral equations 1 into ODEs, we will transform the integral equation describing the larvae population into an equivalent set of differential equations. Assuming $\mathcal{P}_{l,p}(u,t)$ and $\mathcal{P}_{l,d}(u,t)$ are defined in terms of phase-type distribution survival functions,

$$\begin{aligned} \mathcal{P}_{l,p}(u,t) &= \mathcal{S}_{l,p}(\mathcal{D}_{l,p}) = \boldsymbol{\alpha}_{l,p}\ e^{\mathcal{D}_{l,p}\boldsymbol{\mathcal{Q}}_{l,p}}\ \boldsymbol{1}, \\ \mathcal{P}_{l,d}(u,t) &= \mathcal{S}_{l,d}(\mathcal{A}_{l,d}) = \boldsymbol{\alpha}_{l,d}\ e^{\mathcal{A}_{l,d}\boldsymbol{\mathcal{Q}}_{l,d}}\ \boldsymbol{1}, \end{aligned} \tag{7}$$

we can write $l(t)$ as the outcome of the matrix multiplication between a row vector $\boldsymbol{l}$ and an all-ones column vector $\boldsymbol{1}$ as follows,

$$\begin{aligned} &l(t) = \boldsymbol{l}(t)\ \boldsymbol{1}, \\ &\boldsymbol{l}(t) = \int_0^t \lambda_{e,l}(u) \left(\boldsymbol{\alpha}_{l,p}\ e^{\mathcal{D}_{l,p}(u,t)\boldsymbol{\mathcal{Q}}_{l,p}}\right) \otimes \left(\boldsymbol{\alpha}_{l,d}\ e^{\mathcal{A}_{l,d}(u,t)\boldsymbol{\mathcal{Q}}_{l,d}}\right)\ du + l(0) \left(\boldsymbol{\alpha}_{l,p}\ e^{\mathcal{D}_{l,p}(0,t)\boldsymbol{\mathcal{Q}}_{l,p}}\right) \otimes \left(\boldsymbol{\alpha}_{l,d}\ e^{\mathcal{A}_{l,d}(0,t)\boldsymbol{\mathcal{Q}}_{l,d}}\right). \end{aligned} \tag{8}$$

We have expressed $\boldsymbol{l}(t)$ using the Kronecker product of row vectors, denoted by the symbol $\otimes$ in the equation above. Given the definitions of $\mathcal{D}_{l,d}(u,t)$ and $\mathcal{A}_{l,d}(u,t)$ in equations 2 and 4 respectively, it is straight forward to calculate the derivative of $\boldsymbol{l}(t)$ with respect to time,

$$\frac{d\boldsymbol{l}(t)}{dt} = \lambda_{e,l}(t)\ \boldsymbol{\alpha}_{l,p} \otimes \boldsymbol{\alpha}_{l,d} + \boldsymbol{l}(t)\ (\gamma_{l,p}(t)\boldsymbol{\mathcal{Q}}_{l,p} \otimes \boldsymbol{I}_{l,d} + \gamma_{l,d}(t)\boldsymbol{I}_{l,p} \otimes \boldsymbol{\mathcal{Q}}_{l,d}) \tag{9}$$

Where $\boldsymbol{I}_{l,p}$ and $\boldsymbol{I}_{l,d}$ are identity matrices that have dimensions that match those of $\boldsymbol{\mathcal{Q}}_{l,p}$ and $\boldsymbol{\mathcal{Q}}_{l,d}$, respectively. In calculating the derivative of $\boldsymbol{l}(t)$, the following equations were employed.

$$\begin{aligned} \frac{\partial\ e^{\mathcal{D}_{l,p}(u,t)\boldsymbol{\mathcal{Q}}_{l,p}}}{\partial t} &= \gamma_{l,p}(t)\ e^{\mathcal{D}_{l,p}(u,t)\boldsymbol{\mathcal{Q}}_{l,p}}\boldsymbol{\mathcal{Q}}_{l,p} \\ \frac{\partial\ e^{\mathcal{A}_{l,d}(u,t)\boldsymbol{\mathcal{Q}}_{l,d}}}{\partial t} &= \gamma_{l,d}(t)\ e^{\mathcal{A}_{l,d}(u,t)\boldsymbol{\mathcal{Q}}_{l,d}}\boldsymbol{\mathcal{Q}}_{l,d} \end{aligned} \tag{10}$$

Differential equation 9 describes the temporal evolution of the vector $\boldsymbol{l(t)}$, and the larvae population is the sum of the elements in $\boldsymbol{l(t)}$. Similar procedure can be used to convert the integral equations that describe the pupa, egg and adult mosquito populations into differential equations governing the corresponding vectors $\boldsymbol{p(t)}$, $\boldsymbol{e(t)}$, $\boldsymbol{a(t)}$. Furthermore, the emergence rates, which are given in equations 6, could be expressed in terms of the components of the vectors $\boldsymbol{l(t)}$, $\boldsymbol{p(t)}$, $\boldsymbol{e(t)}$, $\boldsymbol{a(t)}$; for instance,

$$\lambda_{l,p}(t) = -\gamma_{l,p}(t)\ \boldsymbol{l}(t)\ (\boldsymbol{\mathcal{Q}}_{l,p} \otimes \boldsymbol{I}_{l,d})\ \boldsymbol{1} = \gamma_{l,p}(t)\ \boldsymbol{l}(t)\ (\boldsymbol{q}_{l,p} \otimes \boldsymbol{1}),$$

where $\boldsymbol{q}_{l,p} = -\boldsymbol{\mathcal{Q}}_{l,p}\ \boldsymbol{1}$.

The components of vectors $\boldsymbol{l(t)}$, $\boldsymbol{p(t)}$, $\boldsymbol{e(t)}$ and $\boldsymbol{a(t)}$ can be rearranged to provide a more intuitive picture of the ODEs describing the mosquito life-cycle. To this end, assume that $\boldsymbol{\mathcal{Q}}_{l,p}$ and $\boldsymbol{\mathcal{Q}}_{l,d}$ in equation 9, are $m_1 \times m_1$ and $m_2 \times m_2$ matrices, respectively, indicating that the corresponding phase-type distributions have m_1 and m_2 transient states. If we rearrange the row vector $\boldsymbol{l}(t)$ into a $m_1 \times m_2$ matrix $\boldsymbol{L}(t)$ defined as:

$$\boldsymbol{L}(t)_{j,s} = \boldsymbol{l}(t)_{(j-1)\times m_2 + s},\ j \in \{1, \cdots, m_1\},\ s \in \{1, \cdots, m_2\},$$

equation 9 can be rewritten as

$$\frac{d\boldsymbol{L}_{j,s}}{dt} = \lambda_{e,l}(t)\ (\boldsymbol{\alpha}_{l,p})_j\ (\boldsymbol{\alpha}_{l,d})_s + \gamma_{l,p}(t) \sum_{k=1}^{r_1} (\boldsymbol{\mathcal{Q}}_{l,p})_{k,j} \boldsymbol{L}_{k,s} + \gamma_{l,d}(t) \sum_{k=1}^{r_2} (\boldsymbol{\mathcal{Q}}_{l,d})_{k,s} \boldsymbol{L}_{j,k}, \tag{11}$$

where the larvae population can be expressed as $l(t) = \sum_{j,s}^{m_1,m_2} \boldsymbol{L}(t)_{j,s}$. The definition of $\boldsymbol{L}$ implies that the total larvae population is distributed among $m_1 \times m_2$ compartments and the elements of $\boldsymbol{L}$ represent the compartments' population. Additionally, non-homogeneous Markovian transitions among these compartments are possible and they are governed by the $\boldsymbol{\mathcal{Q}}$ matrices. Figure 3 illustrates the larvae compartments and the transition rates given by equation 11. The pupa and dead larva states are depicted as the absorbing states "p" and "d" in the figure. It is straight forward to show that the pupa emergence rate is

$$\lambda_{l,p}(t) = \gamma_{l,p}(\tau) \sum_{j,s}^{m_1,m_2} \boldsymbol{L}(t)_{j,s}\ (\boldsymbol{q}_{l,p})_j.$$

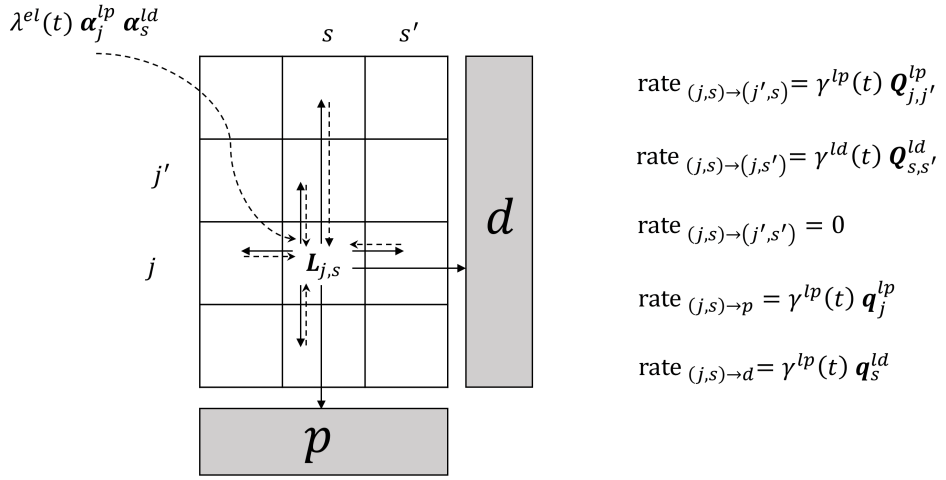


Fig. 3. The diagram illustrates Equation 11, which describes the transition from the larva to pupa and dead larva states, represented as the absorbing states 'p' and 'd,' respectively. By utilizing phase-type distributions for the transition processes in the larva state, we effectively partition the larva state into several auxiliary compartments that represent the history of the original processes. The figure illustrates these larva compartments and the corresponding transition rates as given by Equation 11.

Similar interpretation of the ODEs applies to the other states in the mosquito life-cycle, meaning that, each state divides into several sub-states representing different levels of developments.

The simplest phase-type distribution consists of only one state, and the initial distribution vector and the transition rate matrix are

$$\boldsymbol{\mathcal{Q}} = (-1)\,,\ \boldsymbol{\alpha} = (1)\,.$$

This trivial phase-type distribution is an exponential distribution with mean one. If we assume all the probability functions $\mathcal{P}$ in the integral equations 1 are expressed in terms of the survival function of this particular phase-type distribution, then the corresponding system of ODEs becomes,

$$\begin{aligned}
\frac{dl(t)}{dt} &= \left(1 - \frac{l(t)}{C(t)}\right)\gamma_{e,l}(t)e(t)\ - \gamma_{l,p}(t)\ l(t) - \gamma_{l,d}(t)\ l(t)\\
\frac{dp(t)}{dt} &= \gamma_{l,p}(t)\ l(t)\ - \gamma_{p,a}(t)\ p(t) - \gamma_{p,d}(t)\ p(t)\\
\frac{da(t)}{dt} &= \gamma_{p,a}(t)\ p(t) - \gamma_{a,d}(t)\ a(t)\\
\frac{de(t)}{dt} &= ov(t)\ \gamma_{a,e}(t)\ a(t)\ - \gamma_{e,l}(t)\ e(t) - \gamma_{e,d}(t)\ e(t)
\end{aligned} \tag{12}$$

The system of ODEs above is often used in mechanistic models to calculate mosquito populations. However, these equations allow transitions from a life-cycle stage even if the development in that stage is not complete. This is due to the form of exponential distribution survival function which is depicted in figure 2. A better choice of survival functions that prevent immature transitions is portrayed in the same figure by the black curve which is the Erlang distribution survival function. The Erlang distribution with shape parameter J and mean one can be viewed as a phase-type distribution with J transient states. The initial distribution vector and the transition rate matrix that are as follows

$$\boldsymbol{\alpha} = \begin{pmatrix}1 & 0 & 0 & \cdots & 0 & 0\end{pmatrix}_{1\times J},\ \boldsymbol{\mathcal{Q}} = \begin{pmatrix}
-J & J & 0 & \cdots & 0 & 0\\
0 & -J & J & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & -J & J\\
0 & 0 & 0 & \cdots & 0 & -J
\end{pmatrix}_{J\times J} \tag{13}$$

Assuming that the development probabilities $\mathcal{P}_{l,p}$, $\mathcal{P}_{p,a}$, $\mathcal{P}_{a,e}$ and $\mathcal{P}_{e,l}$ in the integral equations 1 are defined using the survival function of an Erlang distribution with mean one and shape parameter J, and that the mortality probabilities $\mathcal{P}_{.,d}$ are expressed in terms of the survival function an exponential distribution with mean one, the resulting system of ODEs is given by

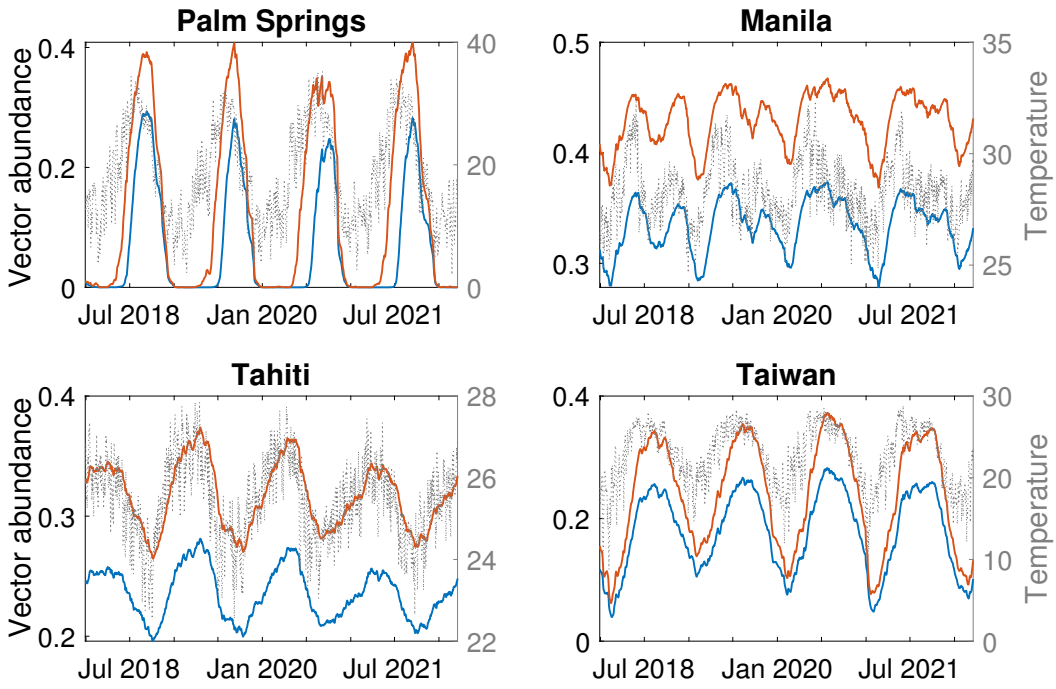


Fig. 4. The figures show simulated vector abundance (arbitrary units) using a non-Markovian model (blue curves) and a Markovian model (red curves). In these simulations, we utilize the daily average temperature for each location indicated in the figure, and the temperature profiles are depicted using gray curves.

$$
\begin{aligned}
\frac{1}{J}\frac{d\boldsymbol{L}_j}{dt} &= \left(1-\frac{l(t)}{C(t)}\right)\gamma_{e,l}(t)\boldsymbol{E}_J\ 1_{j=1} + \gamma_{l,p}(t)\left(1_{j>1}\ \boldsymbol{L}_{j-1} - \boldsymbol{L}_j\right) - \frac{1}{J}\gamma_{l,d}(t)\boldsymbol{L}_j \\
\frac{1}{J}\frac{d\boldsymbol{P}_j}{dt} &= \gamma_{l,p}(t)\boldsymbol{L}_J\ 1_{j=1} + \gamma_{p,a}(t)\left(1_{j>1}\ \boldsymbol{P}_{j-1} - \boldsymbol{P}_j\right) - \frac{1}{J}\gamma_{p,d}(t)\boldsymbol{P}_j \\
\frac{1}{J}\frac{d\boldsymbol{A}_j}{dt} &= \left(\frac{1}{2}\gamma_{p,a}(t)\boldsymbol{P}_J + \gamma_{a,e}(t)\boldsymbol{A}_J\right)\ 1_{j=1} + \gamma_{a,e}(t)\left(1_{j>1}\ \boldsymbol{A}_{j-1} - \boldsymbol{A}_j\right) - \frac{1}{J}\gamma_{a,d}(t)\boldsymbol{A}_j \\
\frac{1}{J}\frac{d\boldsymbol{E}_j}{dt} &= ov(t)\gamma_{a,e}(t)\boldsymbol{A}_J\ 1_{j=1} + \gamma_{e,l}(t)\left(1_{j>1}\ \boldsymbol{E}_{j-1} - \boldsymbol{E}_j\right) - \frac{1}{J}\gamma_{e,d}(t)\boldsymbol{E}_j
\end{aligned}
\tag{14}
$$

where $j \in \{1, \cdots, J\}$, and $1_{j=1}$ is an indicator function. This function is equal to zero for all values of j except when j is equal to one, in which case it takes the value of one. Employing the phase-type distribution with J transient states implies the assumption that the mosquito population in each life-cycle stage is distributed across J sub-states representing different levels of development. Thus the populations of different life-cycle stages are written as

$$
l(t) = \sum_{j=1}^{J}\boldsymbol{L}(t)_j,\ p(t) = \sum_{j=1}^{J}\boldsymbol{P}(t)_j,\ a(t) = \sum_{j=1}^{J}\boldsymbol{A}(t)_j,\ e(t) = \sum_{j=1}^{J}\boldsymbol{E}(t)_j \tag{15}
$$

In order to demonstrate the impact of the distribution of development levels leading to a new life stage on the population of adult mosquitoes, we solved the equations 14 for temperature profiles in four different locations twice: once with the assumption of $J = 1$, which corresponds to exponential distribution, and once with the assumption of $J = 100$, which only allows the transitions with the development levels around 100 percent. Figure 4 shows the results of these calculations. We can observe using a value of $J = 1$ for regions such as Manila and Tahiti, where the temperature varies between 23 and 32 degrees Celsius throughout a year, leads to a higher abundance of the adult mosquitoes compared to using a value of $J = 100$. However, both curves have a similar profile over time, except that the curve for the value of $J = 1$ is shifted upward. We can also note that in regions like Palm Springs, California, where the temperature stays below 15 degrees Celsius for an extended period each year, the difference between the curves for $J = 1$ and $J = 100$ is more distinct. Specifically, the difference between the maximum and minimum abundance on the curve corresponding to $J = 100$ is smaller than that of $J = 1$, and for $J = 1$, the maximum is reached at a later time compared to the case of $J = 100$.

In figure 5, we have compared the time series of mosquito counts obtained from trapping at four distinct sites in Palm Springs[] with the solutions of equations 14. The stars in the plots are the moving averages of the trap counts, calculated

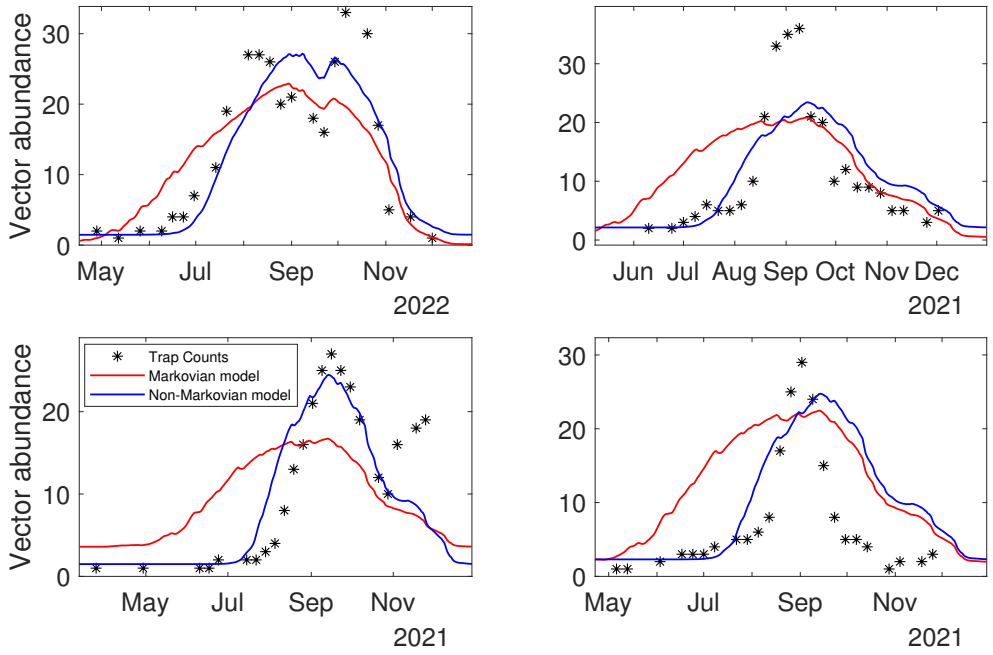


Fig. 5. Comparing the Aedes aegypti trap counts in four different locations in Coachella Valley, California, with simulated vector abundance using a non-Markovian model (blue curves) and a Markovian model (red curves).

as $m_i = (c_{i-1} + c_i + c_{i+1})/3$, where c_i is the trap counts. If a_i^1 represents the calculated adult mosquito population with $J = 1$ and for the time of observation i, we assumed the observations at site s have Poisson distributions with the means $k_s^1 a_i^1 + r_s^1$, where $r_s^1 < 2$. Using an MCMC method we find the values of k_s^1 and r_s^1 that best explain the observations for site s. The red curves in the plots show the fitted $k_s^1 a_i^1 + r_s^1$ curves. We also solved equations 14 with $J = 100$, and the blue curves depict these solutions fitted to observations. We note that for the traps shown in figure 5, the model with $J = 100$ is a better fit to the observed data compared to the model with $J = 1$. This is due to the fact that the model with $J = 100$ more accurately accounts for the longer development time of adult mosquito from egg during the cooler periods of the year. It is important to note that the trap counts within a specific area strongly depend on local factors such as available habitats, temperature condition within the habitats, and other factors that can impact the trapping process, such as wind pattern. Therefore, fitting mechanistic models to the trap counts without a comprehensive understanding of these local factors may not provide a conclusive support for adopting a specific mechanistic model.

Although the carrying capacity, development rates, and mortality rates in equations 14 are time-dependent, it is constructive to examine the steady state of these equations assuming constant rates and constant carrying capacity. By setting the time derivatives on the left-hand-side of the equations to zero, we can find the steady state population of the eggs, larvae, pupae and adult mosquitoes. For the non-trivial equilibrium point, we have determined that the total larvae population is given by

$$\bar{l} = C\left(1 - \frac{1}{r_0}\right),$$

where the basic offspring number r_0 can be computed using the following equation:

$$r_0 = \frac{ov}{\left(1 + \frac{\gamma_{a,d}}{J\gamma_{a,e}}\right)^J - 1}\left(1 + \frac{\gamma_{e,d}}{J\gamma_{e,l}}\right)^{-J}\left(1 + \frac{\gamma_{l,d}}{J\gamma_{l,p}}\right)^{-J}\left(1 + \frac{\gamma_{p,d}}{J\gamma_{p,a}}\right)^{-J}. \tag{16}$$

We have also found that the populations of eggs, pupae and adult mosquitoes are proportional to the larvae population. Therefore, a meaningful non-trivial equilibrium point can only exist if the value of r_0 is greater than 1, and we can show that this equilibrium point is stable for such values of r_0. In fact, the value of r_0 is the average number of viable adult mosquitoes produced by one adult mosquito. To understand the biological interpretation of r_0, note that the first term on the r.h.s of equation 16 represents the expected number of eggs laid by one adult mosquito during its lifespan, and the subsequent terms provides the probability of an egg surviving to the adult form. As a result, the stability condition, $r_0 > 1$, implies that the mosquito population can only survive if each mosquito has more than one viable offspring on average. Furthermore, as $f(x) = (1 + x/J)^J$ is an increasing function of J for any positive x, we can deduce that the value of offspring number r_0 is maximum when $J = 1$. This explains why models with $J > 1$ result in lower mosquito abundance, as depicted in figure 4.

2.2 Dengue transmission model

In the following, we modify the mosquito population model to describe dengue transmission between humans and mosquitoes. Since the mosquito population may vary significantly during the year due to environmental factors such as temperature and habitat changes, any model that aims to describe the dynamics of dengue transmission must account for the mosquito population variation. To this end, we divide the adult female mosquitoes into susceptible, exposed and infectious populations and assume all newly emerged mosquitoes are susceptible and the infectious mosquitoes remain infectious until they die. During its Gonotrophic cycle an infectious mosquito gives a number of bites to humans that can transmit the virus to a susceptible host. Similarly, a susceptible mosquito may acquire the virus through biting infectious humans. Since the extrinsic incubation period of dengue virus in mosquitoes depends on temperature, infected mosquitoes become infectious with a temperature dependent rate. In contrast, we assume that the intrinsic incubation and infectious periods of dengue in human do not significantly vary with temperature. Based on these considerations we can modify integral equations 1 and add the human states in order to describe the transmission dynamics. However, we limit the probability functions in the integral equations to the Erlang distribution, as described in section 2.1. Therefore, we start from equations 14 and incorporate the transmission dynamics by dividing the adult mosquito compartments $\boldsymbol{A}_j$ into susceptible, $\boldsymbol{A}_j^{\boldsymbol{s}}$, exposed, $\boldsymbol{A}_j^{\boldsymbol{e}}$ and infectious , $\boldsymbol{A}_j^{\boldsymbol{i}}$. Furthermore, we can divide the $\boldsymbol{A}_j^{\boldsymbol{e}}$ compartments into additional compartments to account for non-exponential transition from expose state to infectious state. But, we avoid these additional states to keep the presentation of equations simple. As a result, the third equations in set of equations 14 is replaced by the following equations

$$\begin{aligned}
\frac{1}{J}\frac{d\boldsymbol{A}_j^{\boldsymbol{s}}}{dt} &= (\gamma_{p,a}(t)\boldsymbol{P}_J + \gamma_{a,e}(t)\boldsymbol{A}_J^{\boldsymbol{s}})\ 1_{j=1} + \gamma_{a,e}(t)\left(1_{j>1}\ \boldsymbol{A}_{j-1}^{\boldsymbol{s}} - \boldsymbol{A}_j^{\boldsymbol{s}}\right) - \frac{1}{J}\gamma_{a,d}(t)\boldsymbol{A}_j^{\boldsymbol{s}} - \lambda_{HV}(t)\boldsymbol{A}_j^{\boldsymbol{s}} \\
\frac{1}{J}\frac{d\boldsymbol{A}_j^{\boldsymbol{e}}}{dt} &= (\gamma_{p,a}(t)\boldsymbol{P}_J + \gamma_{a,e}(t)\boldsymbol{A}_J^{\boldsymbol{e}})\ 1_{j=1} + \gamma_{a,e}(t)\left(1_{j>1}\ \boldsymbol{A}_{j-1}^{\boldsymbol{e}} - \boldsymbol{A}_j^{\boldsymbol{e}}\right) - \frac{1}{J}\gamma_{a,d}(t)\boldsymbol{A}_j^{\boldsymbol{e}} + \lambda_{HV}(t)\boldsymbol{A}_j^{\boldsymbol{s}} - \gamma_V(t)\boldsymbol{A}_j^{\boldsymbol{e}} \\
\frac{1}{J}\frac{d\boldsymbol{A}_j^{\boldsymbol{i}}}{dt} &= \left(\gamma_{p,a}(t)\boldsymbol{P}_J + \gamma_{a,e}(t)\boldsymbol{A}_J^{\boldsymbol{i}}\right)\ 1_{j=1} + \gamma_{a,e}(t)\left(1_{j>1}\ \boldsymbol{A}_{j-1}^{\boldsymbol{i}} - \boldsymbol{A}_j^{\boldsymbol{i}}\right) - \frac{1}{J}\gamma_{a,d}(t)\boldsymbol{A}_j^{\boldsymbol{i}} + \gamma_V(t)\boldsymbol{A}_j^{\boldsymbol{e}}
\end{aligned} \tag{17}$$

where λ_{HV} is dengue transmission rate from humans to mosquitoes, and γ_V^{-1} is the expected extrinsic period in mosquitoes. We note that the mosquitoes lays eggs regardless of their infection state. Therefore, the first term on the r.h.s of the equation describing the egg populations should be modified as follows

$$\frac{1}{J}\frac{d\boldsymbol{E}_j}{dt} = ov(t)\gamma_{a,e}(t)\left(\boldsymbol{A}_J^{\boldsymbol{s}} + \boldsymbol{A}_J^{\boldsymbol{e}} + \boldsymbol{A}_J^{\boldsymbol{i}}\right)\ 1_{j=1} + \gamma_{e,l}(t)\left(1_{j>1}\ \boldsymbol{E}_{j-1} - \boldsymbol{E}_j\right) - \frac{1}{J}\gamma_{e,d}(t)\boldsymbol{E}_j \tag{18}$$

With regard to humans, we consider four sub-populations representing susceptible, $\boldsymbol{H}^{\boldsymbol{s}}$, exposed, $\boldsymbol{H}^{\boldsymbol{e}}$, infectious, $\boldsymbol{H}^{\boldsymbol{i}}$, and recovered, $\boldsymbol{H}^{\boldsymbol{r}}$, human populations. Assuming the transitions between the human states are Markovian we can write

$$\begin{aligned}
\frac{d\boldsymbol{H}^{\boldsymbol{s}}}{dt} &= -\lambda_{VH}(t)\boldsymbol{H}^{\boldsymbol{s}} \\
\frac{d\boldsymbol{H}^{\boldsymbol{e}}}{dt} &= \lambda_{VH}(t)\boldsymbol{H}^{\boldsymbol{s}} - \gamma_H\boldsymbol{H}^{\boldsymbol{e}} \\
\frac{d\boldsymbol{H}^{\boldsymbol{i}}}{dt} &= \gamma_H\boldsymbol{H}^{\boldsymbol{e}} - \eta_H\boldsymbol{H}^{\boldsymbol{i}} \\
\frac{d\boldsymbol{H}^{\boldsymbol{r}}}{dt} &= \eta_H\boldsymbol{H}^{\boldsymbol{i}}
\end{aligned} \tag{19}$$

In the equations above, λ_{VH} is dengue transmission rate from mosquitoes to humans. Also, the constant parameters γ_H^{-1} and η_H^{-1} are, respectively, the expected intrinsic and infectious periods in humans. The infection transmission rate from vector to host can be expressed in terms of average number of bites a vector gives to the hosts in a gonotrophic cycle, n_B, and the expected length of gonotrophic cycle, $\gamma_{a,e}^{-1}$, as follows

$$\lambda_{VH} = n_B\gamma_{a,e}\phi_{VH}\frac{1}{N_H}\sum_j \boldsymbol{A}_j^{\boldsymbol{i}}, \tag{20}$$

where, N_H is the human population and ϕ_{VH} represents the probability of an infected vector transmitting the infection to a host through a single bite. We have assumed that biting occurs at every stage of the gonotrophic cycle. But, if we restrict biting to only the later k stages of the cycle, then in the equation above, the summation should be only on those stages. Moreover, we should multiply the r.h.s of the equation by the ratio of J (the total number of stages in the cycle) to k. If the probability of transmitting the infection from an infected human to a susceptible mosquito via a single bite is represented by ϕ_{HV}, we can write the infection transmission rate from humans to vectors as

$$\lambda_{HV} = n_B\gamma_{a,e}\phi_{HV}\frac{1}{N_H}\boldsymbol{H}^{\boldsymbol{i}}. \tag{21}$$

Figure 6 illustrates the transmission model diagram. In case we consider that the mosquitoes' transition from the exposed state to the infectious state is non-Markovian, we need to divide $\boldsymbol{A}_j^{\boldsymbol{e}}$ compartment into multiple states corresponding to different stages of the extrinsic incubation period. Therefore, the row that represents the mosquito exposed state in

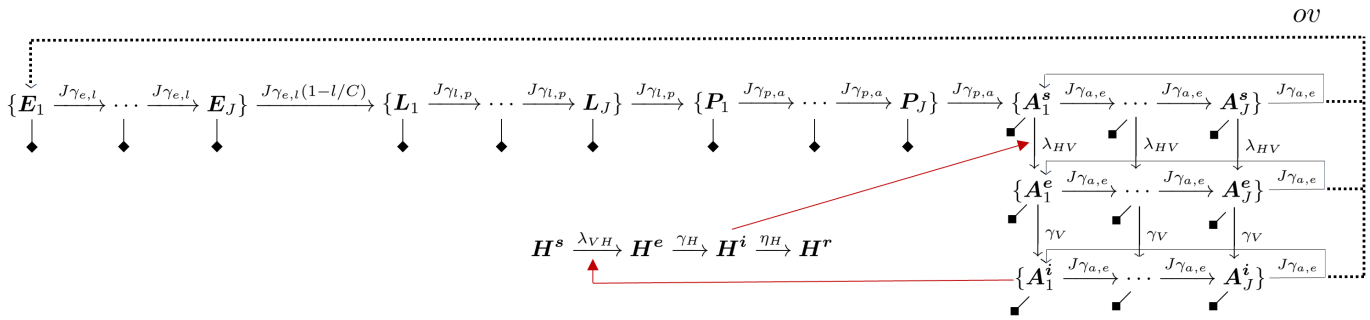


Fig. 6. The diagram depicts a non-Markovian dengue transmission model. The Aedes mosquitoes life cycle, is modeled using the Erlang distribution with temperature-dependent transition rates. In the diagram, the populations of egg, larva, pupa, and adult mosquitoes are represented by E_k, L_k, P_k, and A_k, respectively, indicating different stages of development denoted by the subscript k. The human population is divided into susceptible individuals (H^s), exposed individuals (H^e), infectious individuals (H^i), and recovered individuals (H^r). Similarly, adult mosquitoes can be susceptible (A^s), exposed (A^e), or infectious (A^i). The red arrows indicate infectious transmission, while the black arrows represent transitions between two stages. Oviposition process is denoted by ov.

the diagram should be replicated. To find the dengue basic reproduction number, R_0, under the assumption of constant temperature, we can analyze the stability of disease-free equilibrium.If we linearize the dengue transmission model around the steady state of mosquito population, using the techniques introduced in [26], we can show that R_0 is

$$R_0 = \frac{\bar{a}}{N_H} n_B^2 \gamma_{a,e}^2 \phi_{HV} \phi_{VH} \gamma_{a,d}^{-1} \eta_H^{-1} \left(1 + \frac{\gamma_{a,d}}{\gamma_V}\right)^{-1}. \tag{22}$$

The value of R_0 can be interpreted as the number female mosquitoes that become infected after introducing one infected female mosquitoes into a susceptible population of vectors and hosts. Although this interpretation is only valid for constant temperatures, it can still serve as an estimate for the degree of virus transmission if the temperature fluctuations are not too significant over a time frame longer than the inverse of the transition rates in the transmission model.

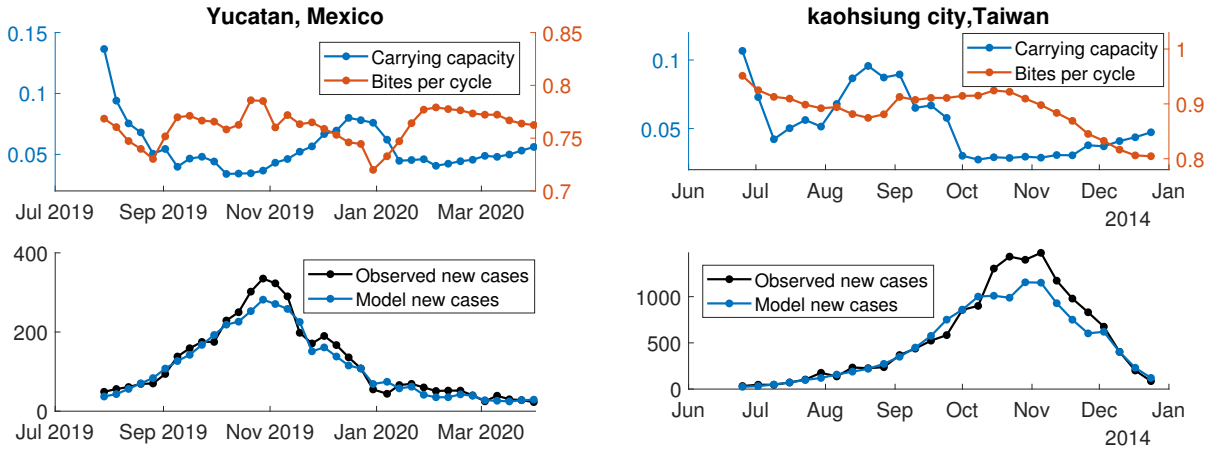
3 NUMERICAL RESULTS

Here, we apply the equations we developed in the previous section to construct a global risk map for dengue. This risk map is based on the entomological characteristics of Aedes aegypti mosquitoes, as it is the main vector for dengue. To develop the risk map, we first estimate the environmental carrying capacity and biting rate for the mosquitoes. Using these estimated parameters and the other model parameters available in related literature, we can calculate the dengue reproduction number and the vector abundance. By correlating these variables with historical human dengue case data, we build a risk model that uses the reproduction number and vector abundance as predictive variables. This enables us to create a global spatio-temporal dengue risk map.

3.1 Estimating model parameters using particle filter

With the exception of the carrying capacity and the number of bites per gonotrophic cycle, all parameters related to dengue transmission model in section 2.2 can be found in literature [17], [22], [23], [7], [9], [14], [24], [2]. The carrying capacity, in general, is a variable that can be impacted by multiple factors, including precipitation and the abundance of water containers within the mosquito habitat. Therefore, constructing a comprehensive model that relates carrying capacity to an assumed set of driving variables requires numerous observations in different locations to cover the variables space. However, data on mosquito abundance are often limited and primarily based on mosquito trap data, which may not accurately reflect the exact number of mosquitoes in a given location. In contrast, dengue transmission data is more readily available, but it is often presented as aggregated numbers of new cases over a wide geographical area. Here, we use dengue new cases time series from 61 different locations [27], [28], [29], [30], [31], [32], [33], [34], to estimate the carrying capacity and number of bites per gonotrophic cycle. To achieve this, we employ the particle filter algorithm to estimate the unknown transmission model parameters.

let $X_k = (V_k, H_k, C_k, n_{Bk})$ represent the transmission system at time step k, which includes the state vectors for mosquitoes and humans denoted as V_k and H_k, as well as the carrying capacity C_k and the number of bites per gonotrophic cycle n_{Bk}. To determine the system state at time step $k+1$, we first sample C_{k+1} and n_{Bk+1} from predefined probability distributions. Next, we evolve the mosquitoes and humans vector using the transmission model, taking into account the updated values of the carrying capacity and number of bites per gonotrophic cycle, C_{k+1} and n_{Bk+1}. This enables us to obtain a sample for the transmission system at time step $k+1$, denoted as $X_{k+1} = (V_{k+1}, H_{k+1}, C_{k+1}, n_{Bk+1})$. To evaluate the validity of a transmission system sample for time step $k+1$, we compare the number of new human dengue cases derived from X_{k+1} with the observed new cases. In doing so, we assume that the discrepancy between the two



(a) (b)

Fig. 7. This figure shows the results of fitting observed human dengue cases to the non-Markovian transmission model described in Section 2.2 for two different locations. The unknown model parameters, which are the carrying capacity and the number of bites per gonotrophic cycle, were estimated using the particle filter algorithm for each location (as shown in the plots in the first row). The plots in the second row demonstrate the performance of fitting the model by comparing the observed human cases with the corresponding values obtained from fitting the transmission model.

values follows a predefined probability density function that peaks at zero. Consequently, we can assign a weight to each sample, which reflects its validity. If we draw a set of such samples, we obtain a vector of weights that we normalize it and thereby assign a probability to the each sample in the set. By selecting the sample set elements according to their assigned probabilities, we obtain an updated set of samples that represents the distribution for the transmission system at time step $k+1$. The particle filtering procedure we described here provides an approximate numerical distribution for the system state, considering a series of observations, i.e., $p(X_{k+1}|n_{k+1}, n_k, \cdots, n_1)$, where n_i is the observed new cases at time step i. However, in our calculation, we adopt the results of backward particle filter smoothing, which provide a distribution for the trajectory of system state, formally denoted as $p(X_{k+1}, X_k, \cdots, X_1|n_{k+1}, n_k, \cdots, n_1)$. In figure 7 we have shown the estimated marginal distribution means for the carrying capacity per human, number of bites per gonotrophic cycle and the fitted number of human new dengue cases for two different locations. For these estimations, in the particle filter algorithm, we draw the value of the carrying capacity at step $k+1$, denoted as C_{k+1}, from a normal distribution. The mean and standard deviation of this distribution are equal to the carrying capacity at step k, denoted as C_k, and 50 percent of C_k, respectively. We draw the value of n_{Bk+1} in a similar manner except we use 5 percent of n_{Bk} as the standard deviation of the normal distribution. We opt to draw the carrying capacity value from a broad distribution due to its susceptibility to significant impacts from factors such as rainfall and other environmental conditions.

To investigate the relationship between carrying capacity and precipitation, we fitted the new human case numbers from 61 locations and generated estimated values similar to those shown in Figure 7. In the left panel of Figure 11d, we plotted the estimated carrying capacity values versus the 2-week moving average precipitation. To provide further insight, the right panel of Figure 11d displays the average values of the estimated carrying capacity within precipitation bins of width 0.002 m, indicated by a cross marker. This plot illustrates that the expected value of carrying capacity increases with low precipitation amounts, but it decreases with large amounts of precipitation. Additionally, we performed fittings of different distributions to the carrying capacity data within each precipitation bin, and we observed that the inverse Gaussian distribution provided a better fit to the data. The inverse Gaussian distribution is a distribution with two parameters that has support on the positive real numbers $(0, \infty)$. If a random variable C follows an inverse Gaussian distribution, it is denoted as $C \sim IG(\mu, \lambda)$, where μ and λ are the distribution parameters. The probability density function for C can be expressed as:

$$f(c|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi c^3}} \exp\left[-\frac{\lambda(c-\mu)^2}{2\mu^2 c}\right].$$

To fit a statistical model to the carrying capacity data presented in the left panel of Figure 11d, we made the assumption that the carrying capacity at each precipitation value follows an inverse Gaussian distribution with the parameters μ and λ that are continuous functions of precipitation. Since the expected value of carrying capacity has different trends for low and high levels of precipitation, we considered the following functions for the model parameters:

$$\mu(p) = \begin{cases} a_0 + a_1 p + a_2 p^2 & \text{if } p < p_0, \\ \alpha_0 + e^{-\alpha_1 p} + \alpha_2 & \text{if } p \geq p_0, \end{cases}$$

and

$$\lambda(p) = \begin{cases} b_0 + b_1 p + b_2 p^2 & \text{if } p < p_0, \\ \beta_0 + \beta_1 p + \beta_2 p^2 & \text{if } p \geq p_0. \end{cases}$$

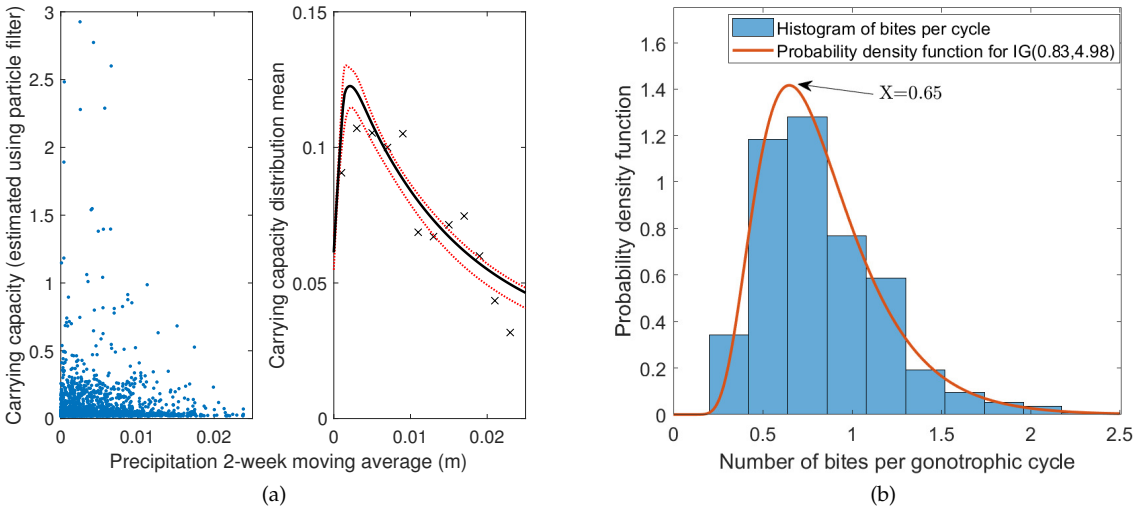


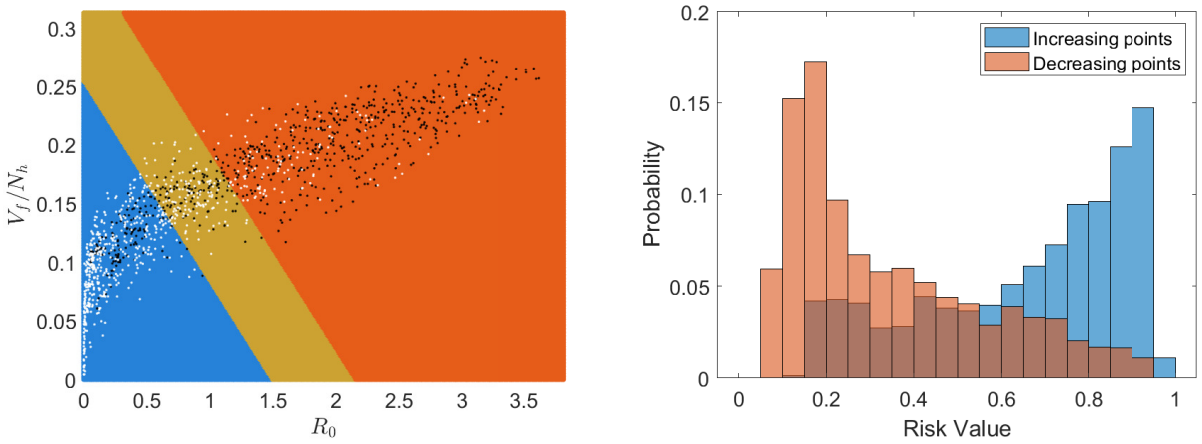
Fig. 8. The left panel of figure (a) shows the estimated carrying capacity values versus the 2-week moving average precipitation. Cross markers in the right panel of figure (a) display the average values of the estimated carrying capacity within precipitation bins of width 0.002 m, and the solid black curve shows the estimated mean of the inverse Gaussian model fitted to the carrying capacity values in the left panel. The dotted red curves represent the 25th and 75th percentiles of the estimated mean. In plot (b), we have shown the histogram of all the estimated values for bites per gonotrophic cycle, with the density function for the fitted inverse Gaussian distribution shown by the red curve.

Here, p represents precipitation, and p_0 is a threshold value. Although the model appears to have 13 parameters, namely $a_0, a_1, a_2, \alpha_0, \alpha_1, \alpha_2, b_0, b_1, b_2, \beta_0, \beta_1, \beta_2$, and p_0, we can express two of the parameters, such as α_0 and β_0, in terms of the remaining parameters due to the continuity of the functions at p_0. Using Gibbs sampling we fitted the the statistical model explained above to the carrying capacity data shown in the left panel of figure 11d. Since we utilized Bayesian inference to fit the model, we are able to derive marginal posterior distributions for the parameters μ and λ at each precipitation value. In the right panel of Figure 11d, the black solid line represents the mean of the marginal posterior distribution of μ, while the dotted red curve represents the 25th and 75th percentiles. For a random variable with an inverse Gaussian distribution $IG(\mu, \lambda)$, the expected value is equal to the parameter μ. Therefore, the black solid line in the figure represents the expected value of the carrying capacity. Although we might expect that increasing precipitation leads to an elevated carrying capacity, our results show that this is not always true. However, further data collection is needed to assert this relationship with a high certainty. One explanation for the lower carrying capacity in high precipitation environment might be due to the fact that Aedes aegypti is an urban mosquito, and their breeding sites may become flooded during periods of very high precipitation.

Fitting the new human dengue cases in 61 locations to the spreading model also provided estimated values for the number of bites per gonotrophic cycle, similar to those shown in figures 7. A histogram of all the estimated values for bites per gonotrophic cycle is presented in figure 8b. Additionally, we found that the inverse Gaussian distribution with the mean of 0.83 provides the best fit to the estimated values. Although this value is close to one blood meal per cycle, the lower expected value suggests that not all blood meals are successful in the transmission of the dengue virus.

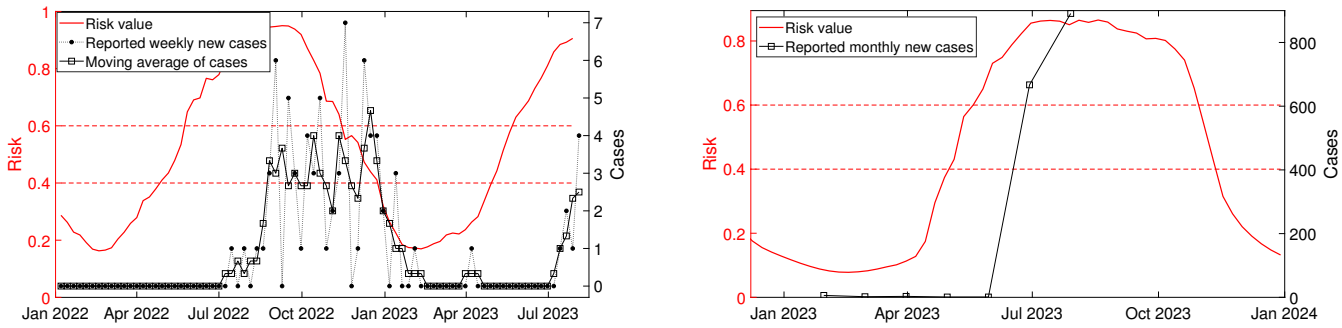
3.2 Dengue risk map

By definition, the dengue reproduction number, given by the equation 22, provides a threshold that determines whether the virus can spread in a healthy population and result in an increasing number of new cases or not. While some studies [2] have utilized the value of R_0 as a risk factor, the application of this parameter necessitates certain considerations. First, the reproduction number is derived under the assumption of a constant temperature, providing an epidemic threshold based on this specific condition. Second, the calculation of R_0 requires an estimation of the number of female mosquitoes, along with its variations influenced by precipitation and temperature. To address these issues, we only use reproduction number as a predictive variable, coupled with an estimation of the mosquito population within a numerical model for assessing the dengue risk factor. Furthermore, we incorporate the model developed for the fluctuation in carrying capacity, as outlined in section 3.1, to estimate the variation in mosquito population. This variation plays a pivotal role in the reproduction number given by equation 22. To create a numerical model for the dengue transmission risk factor, we initially manually labeled the dengue epidemic curves from various locations, [27], [28], [29], [30], [31], [32], [33], [34], with either a one or zero, indicating the weeks with increasing or decreasing cases. Subsequently, we calculated the corresponding values of the reproduction number and mosquito population for the labeled weeks. This calculation relied on temperature and precipitation data for the respective locations, which we obtained from the ERA5 global reanalysis data [35]. ERA5 provides daily data with a resolution of 0.25 degrees in longitude and latitude and is accessible through the Copernicus Climate Change Service [36]. In figure 9a, we have shown random samples of the weeks labeled with zero and one as white and black dots, respectively. Furthermore the x and y values of a labeled week in the plot correspond to the 30-day moving average of the estimated



(a) (b)

Fig. 9. Plot(a) displays random samples of the weeks with increasing and decreasing dengue cases in different locations as black and white dots, respectively. The R_0 and V_f values in the plot correspond to the 30-day moving average of the estimated reproduction number and mosquito population for the respective week. Furthermore, in this plot, the region shaded in red has an outbreak probability greater than 0.6 according to the trained model described in Section 3.2. Meanwhile, the outbreak probability in the blue-shaded area is less than 0.4. Plot(b) displays the distribution of calculated risk values for the training data, which includes weeks with both increasing and decreasing human dengue cases.



(a) (b)

Fig. 10. The plots show human new dengue cases along with the calculated transmission risk for Florida, USA (a), and Sunsari, Nepal (b).

reproduction number and mosquitoes population for the corresponding week. To train a numerical model for dengue transmission risk, we fed the labeled data into a feedforward neural network, which assigns a value between zero and one to points on the two-dimensional surface of reproduction number and mosquito population. This value can be interpreted as the probability of dengue outbreak or having increasing number of cases. It is possible to train neural network models with numerous nodes in the layers, effectively dividing the two-dimensional surface into high and low-risk areas in a nonlinear manner. However, to avoid overfitting, we used a neural network with only one node, which effectively corresponds to a logistic regression model. In figure 9a, we have depicted the classification resulting from the trained model. The region with assigned values greater than 0.6 is colored in red, while the region with values smaller than 0.4 is colored in blue. The slope of the boundaries of the regions in this figure indicates that the R_0 value is not the sole factor determining the evaluation of transmission risk. This might be due to the fact that the R_0 expression is derived for constant temperature and includes some terms that can experience significant fluctuations with temperature. To further demonstrate the fitting performance, in figure 9b, we have shown the distribution of the calculated risk values for the training data points. Although the distributions for the increasing and decreasing points are not completely separated, they are concentrated in the regions of high and low-risk values, respectively.

To demonstrate the utility of the risk model, in Figure 10, we have plotted the human new dengue cases along with the calculated transmission risk for Florida, USA (a), and Sunsari, Nepal (b). We can observe that local transmission occurs during high-risk periods. For instance, in Florida, between January and March 2023, when the risk is low, the number of new cases exhibits a declining trend.

Transmission risk values can be determined for any location, provided access to local temperature and precipitation data. To accomplish this, we obtained temperature and precipitation data for the year 2022 from the Copernicus Climate Change Service [36], which offers daily global data with a resolution of 0.25 degrees in both latitude and longitude. By

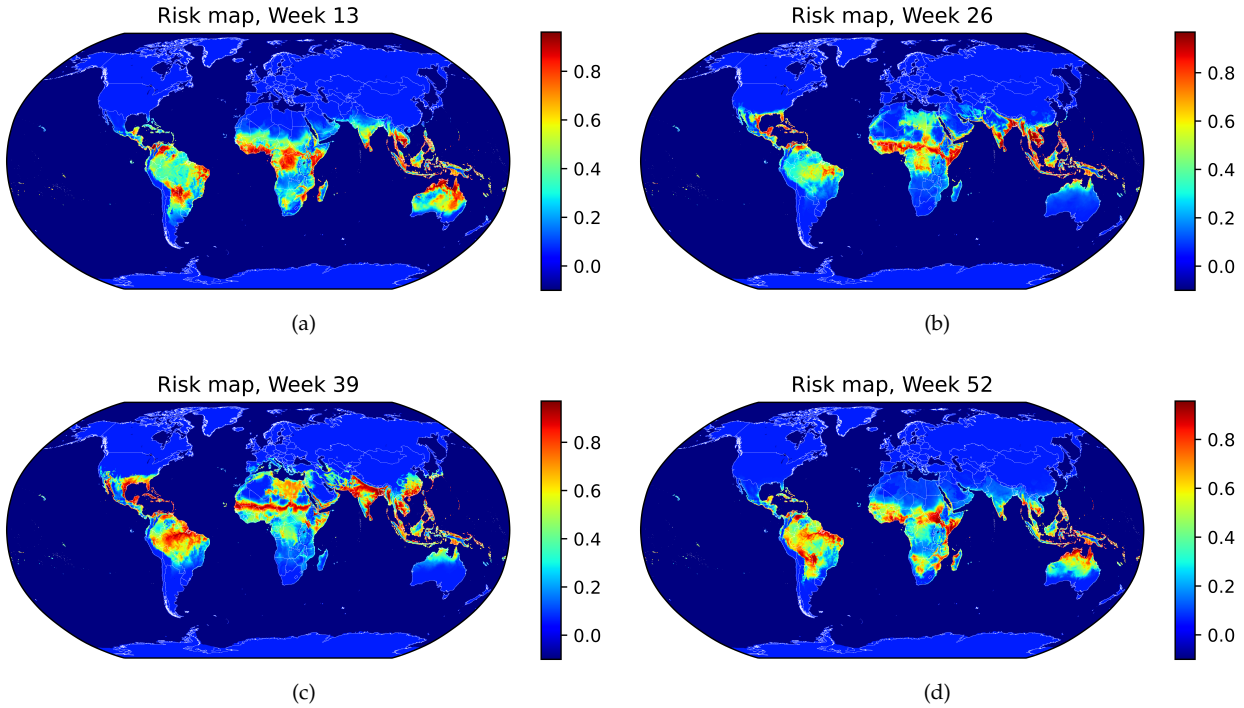


Fig. 11. The figures show global dengue transmission risk for four different weeks in the year 2022.

utilizing the daily average data, we constructed the daily risk map. Figure 11 presents risk maps at four distinct time points throughout the year. When interpreting these maps, it's important to consider that our calculations assume the presence of Aedes aegypti mosquitoes and the dengue virus in a given location, and subsequently assess the likelihood of mosquito survival and dengue transmission. As a result, these maps can indicate the suitability of various locations for the survival of the mosquito population.

4 Conclusions

The main contributions of this study are: a. Developing a new model for the mosquito life cycle that accounts for temperature-dependent development rates. This model improves upon other existing models, which often provide only approximate descriptions of the complex non-Markovian process; b. Creating a global dengue risk map with daily temporal resolution and a spatial resolution of 0.25 degrees in longitude and latitude.

In our model, we assume that the daily development of mosquitoes at each life-cycle phase depends on the average daily temperature. We accumulate daily development values until the mosquitoes' development is complete, at which point they transition to the subsequent phase. Our calculations indicate that in cases of significant temperature fluctuations throughout a year, particularly with prolonged periods of lower temperatures, the traditional non-homogeneous ODEs with temperature-dependent rates predicts an earlier increase in the mosquito population compared to the output of our model. This discrepancy arises from the inherent assumption within traditional non-homogeneous ODEs, as demonstrated in section 2.1, that the transition to a new lifecycle stage occurs when the development reaches a value that is distributed exponentially with a mean of one. This leads to earlier mosquito emergence compared to our model, where the transition occurs upon completion of development. This difference becomes more pronounced in environments with lower temperatures, which result in longer development times.

Although the mosquito life cycle is non-Markovian process and is therefore best described by integral equations, we have shown that it is possible to reduce the master equations to ODEs by assuming the general class of phase-type distribution for the random variables associated with the development values at the transition times. This is particularly useful because it enabled us to analyze the system using well-established methods developed for ODEs and derive the corresponding reproduction number of dengue and the mosquito's offspring number. Furthermore, the reduction to ODEs enabled us to utilize the particle filter algorithm for estimating the environmental carrying capacity and biting rate by fitting the outbreak data to the dengue transmission model. This is significant because the particle filter approach is developed for Markovian systems.

The developed risk map relies on historical outbreak data for the carrying capacity model and outbreak risk assessment itself. While we utilize outbreak data from various global locations, there is room for enhancing both the carrying capacity and risk models. Our model, which links carrying capacity to precipitation, demonstrates a decrease in carrying capacity when the 2-week moving average of daily precipitation exceeds 1 cm. This reduction could be attributed to potential

mosquito breeding site overflow in urban areas. However, further data collection is needed to solidify the relationship between carrying capacity and precipitation. Upon comparing the resultant risk values with actual human dengue case data from Florida and Nepal, we observed a strong correlation between the risk model and the case data. This suggests that the risk assessment tool has practical utility in guiding intervention measures, including mosquito control strategies. Additionally, the risk map assigns a high risk to locations like southern Iran, which have only recently been invaded by Aedes mosquitoes. This observation holds significance because our risk model relies solely on precipitation and temperature data. This indicates that the risk model has the potential to predict areas where Aedes mosquito populations could establish themselves, even in locations where they are not currently present.

ACKNOWLEDGMENTS

The project was sponsored by the Department of the Army, U.S. Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Ft Detrick, MD. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] 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 *et al.*, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[2] E. A. Mordecai, J. M. Cohen, M. V. Evans, P. Gudapati, L. R. Johnson, C. A. Lippi, K. Miazgowicz, C. C. Murdock, J. R. Rohr, S. J. Ryan *et al.*, "Detecting the impact of temperature on transmission of zika, dengue, and chikungunya using mechanistic models," *PLoS neglected tropical diseases*, vol. 11, no. 4, p. e0005568, 2017.

[3] T. W. Scott and W. Takken, "Feeding strategies of anthropophilic mosquitoes result in increased risk of pathogen transmission," *Trends in parasitology*, vol. 28, no. 3, pp. 114–121, 2012.

[4] J. P. Messina, M. U. Kraemer, O. J. Brady, D. M. Pigott, F. M. Shearer, D. J. Weiss, N. Golding, C. W. Ruktanonchai, P. W. Gething, E. Cohn *et al.*, "Mapping global environmental suitability for zika virus," *elife*, vol. 5, p. e15272, 2016.

[5] M. Heydari, M. Metanat, M.-A. Rouzbeh-Far, S. M. Tabatabaei, M. Rakhshani, N. Sepehri-Rad, and M. Keshtkar-Jahromi, "Dengue fever as an emerging infection in southeast iran," *The American Journal of Tropical Medicine and Hygiene*, vol. 98, no. 5, p. 1469, 2018.

[6] S. T. R. d. Pinho, C. P. Ferreira, L. Esteva, F. R. Barreto, V. Morato e Silva, and M. Teixeira, "Modelling the dynamics of dengue real epidemics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 368, no. 1933, pp. 5679–5693, 2010.

[7] H. M. Yang, M. d. L. da Graca Macoris, K. C. Galvani, and M. T. M. Andrighetti, "Follow up estimation of aedes aegypti entomological parameters and mathematical modellings," *Biosystems*, vol. 103, no. 3, pp. 360–371, 2011.

[8] H. Yang, M. Macoris, K. Galvani, M. Andrighetti, and D. Wanderley, "Assessing the effects of temperature on dengue transmission," *Epidemiology & Infection*, vol. 137, no. 8, pp. 1179–1187, 2009.

[9] ——, "Assessing the effects of temperature on the population of aedes aegypti, the vector of dengue," *Epidemiology & Infection*, vol. 137, no. 8, pp. 1188–1202, 2009.

[10] O. J. Brady, N. Golding, D. M. Pigott, M. U. Kraemer, J. P. Messina, R. C. Reiner Jr, T. W. Scott, D. L. Smith, P. W. Gething, and S. I. Hay, "Global temperature constraints on aedes aegypti and ae. albopictus persistence and competence for dengue virus transmission," *Parasites & vectors*, vol. 7, no. 1, pp. 1–17, 2014.

[11] S.-C. Chen and M.-H. Hsieh, "Modeling the transmission dynamics of dengue fever: implications of temperature effects," *Science of the total environment*, vol. 431, pp. 385–391, 2012.

[12] H. M. Yang, "The transovarial transmission in the dynamics of dengue infection: Epidemiological implications and thresholds," *Mathematical biosciences*, vol. 286, pp. 1–15, 2017.

[13] E. P. Pliego, J. Velázquez-Castro, and A. F. Collar, "Seasonality on the life cycle of aedes aegypti mosquito and its statistical relation with dengue outbreaks," *Applied Mathematical Modelling*, vol. 50, pp. 484–496, 2017.

[14] M. M. Rossi, L. Ólivêr, and E. Massad, "Modelling the implications of temperature on the life cycle of aedes aegypti mosquitoes," *Ecological Modelling Applied to Entomology*, pp. 81–107, 2014.

[15] M. Otero, N. Schweigmann, and H. G. Solari, "A stochastic spatial dynamical model for aedes aegypti," *Bulletin of mathematical biology*, vol. 70, pp. 1297–1325, 2008.

[16] M. U. Kraemer, M. E. Sinka, K. A. Duda, A. Q. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. Van Bortel *et al.*, "The global distribution of the arbovirus vectors aedes aegypti and ae. albopictus," *elife*, vol. 4, p. e08347, 2015.

[17] D. A. Focks, D. G. Haile, E. Daniels, and G. A. Mount, "Dynamic life table model for aedes aegypti (diptera: Culicidae): analysis of the literature and model development," *Journal of medical entomology*, vol. 30, no. 6, pp. 1003–1017, 1993.

[18] Z. Feng, K. Gurski, O. Prosper, M. Teboh-Ewungkem, and M. Grogan, "A mosquito-borne disease model with non-exponentially distributed infection and treatment stages," *Journal of Dynamics and Differential Equations*, vol. 33, pp. 1679–1709, 2021.

[19] C. Nowzari, M. Ogura, V. M. Preciado, and G. J. Pappas, "A general class of spreading processes with non-markovian dynamics," in *2015 54th IEEE conference on decision and control (CDC)*. IEEE, 2015, pp. 5073–5078.

[20] R. Forien, G. Pang, and É. Pardoux, "Estimating the state of the covid-19 epidemic in france using a model with memory," *Royal Society open science*, vol. 8, no. 3, p. 202327, 2021.

[21] A. Vajdi, L. W. Cohnstaedt, L. E. Noronha, D. N. Mitzel, W. C. Wilson, and C. M. Scoglio, "A non-markovian model to assess contact tracing for the containment of covid-19," *IEEE Transactions on Network Science and Engineering*, 2023.

[22] L. Eisen, A. J. Monaghan, S. Lozano-Fuentes, D. F. Steinhoff, M. H. Hayden, and P. E. Bieringer, "The impact of temperature on the bionomics of aedes (stegomyia) aegypti, with special reference to the cool geographic range margins," *Journal of medical entomology*, vol. 51, no. 3, pp. 496–516, 2014.

[23] 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, p. e87468, 2014.

[24] L. C. Farnesi, A. J. Martins, D. Valle, and G. L. Rezende, "Embryonic development of aedes aegypti (diptera: Culicidae): influence of different constant temperatures," *Memórias do Instituto Oswaldo Cruz*, vol. 104, pp. 124–126, 2009.

[25] D. R. Cox and H. D. Miller, *The theory of stochastic processes*. CRC press, 1977, vol. 134.

[26] O. Diekmann, J. Heesterbeek, and M. G. Roberts, "The construction of next-generation matrices for compartmental epidemic models," *Journal of the royal society interface*, vol. 7, no. 47, pp. 873–885, 2010.

[27] Paho/who data. Subnational cases from Brazil and Mexico, Accessed on May, 2021. [Online]. Available: https://www3.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en.html
[28] A. Murphy, G. S. Rajahram, J. Jilip, M. Maluda, T. William, W. Hu, S. Reid, G. J. Devine, and F. D. Frentiu, "Incidence and epidemiological features of dengue in sabah, malaysia," *PLoS neglected tropical diseases*, vol. 14, no. 5, p. e0007504, 2020.
[29] M. Chaudhry, S. Ahmad, H. B. Rashid, and I. Uddin, "Dengue epidemic in postconflict swat district, khyber pakhtunkhwa, pakistan, 2013," *The American journal of tropical medicine and hygiene*, vol. 96, no. 4, p. 899, 2017.
[30] Z. Xu, H. Bambrick, L. Yakob, G. Devine, F. D. Frentiu, F. V. Salazar, R. Bonsato, and W. Hu, "High relative humidity might trigger the occurrence of the second seasonal peak of dengue in the philippines," *Science of The Total Environment*, vol. 708, p. 134849, 2020.
[31] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, "Forecast of dengue incidence using temperature and rainfall," *PLoS neglected tropical diseases*, vol. 6, no. 11, p. e1908, 2012.
[32] P. Liyanage, H. Tissera, M. Sewe, M. Quam, A. Amarasinghe, P. Palihawadana, A. Wilder-Smith, V. R. Louis, Y. Tozan, and J. Rocklöv, "A spatial hierarchical analysis of the temporal influences of the el nino-southern oscillation and weather on dengue in kalutara district, sri lanka," *International journal of environmental research and public health*, vol. 13, no. 11, p. 1087, 2016.
[33] P. Vásquez, A. Loría, F. Sanchez, and L. A. Barboza, "Climate-driven statistical models as effective predictors of local dengue incidence in costa rica: A generalized additive model and random forest approach," *Revista de Matematica: Teoría y Aplicaciones*, vol. 27, no. 1, pp. 1–21, 2020.
[34] H. Q. Cuong, N. T. Hien, T. N. Duong, T. V. Phong, N. N. Cam, J. Farrar, V. S. Nam, K. T. Thai, and P. Horby, "Quantifying the emergence of dengue in hanoi, vietnam: 1998–2009," *PLoS neglected tropical diseases*, vol. 5, no. 9, p. e1322, 2011.
[35] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.
[36] Copernicus climate change service. [Online]. Available: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview

Am. J. Trop. Med. Hyg., 104(4), 2021, pp. 1444–1455
doi:10.4269/ajtmh.20-0444
Copyright © 2021 by The American Society of Tropical Medicine and Hygiene

Risk Assessment of Dengue Transmission in Bangladesh Using a Spatiotemporal Network Model and Climate Data

Mahbubul H. Riad,[1] Lee W. Cohnstaedt,[2]* and Caterina M. Scoglio[1]

[1]*Department of Electrical and Computer Engineering, College of Engineering, Kansas State University, Manhattan, Kansas;* [2]*United States Department of Agriculture, Arthropod-borne Animal Diseases Research, Manhattan, Kansas*

Abstract. Vector-borne disease risk assessment is crucial to optimize surveillance, preventative measures (vector control), and resource allocation (medical supplies). High arthropod abundance and host interaction strongly correlate to vector-borne pathogen transmission. Increasing host density and movement increases the possibility of local and long-distance pathogen transmission. Therefore, we developed a risk-assessment framework using climate (average temperature and rainfall) and host demographic (host density and movement) data, particularly suitable for regions with unreported or underreported incidence data. This framework consisted of a spatiotemporal network-based approach coupled with a compartmental disease model and nonhomogeneous Gillespie algorithm. The correlation of climate data with vector abundance and host–vector interactions is expressed as vectorial capacity—a parameter that governs the spreading of infection from an infected host to a susceptible one via vectors. As an example, the framework is applied for dengue in Bangladesh. Vectorial capacity is inferred for each week throughout a year using average monthly temperature and rainfall data. Long-distance pathogen transmission is expressed with human movement data in the spatiotemporal network. We have identified the spatiotemporal suitability of dengue spreading in Bangladesh as well as the significant-incidence window and peak-incidence period. Analysis of yearly dengue data variation suggests the possibility of a significant outbreak with a new serotype introduction. The outcome of the framework comprised spatiotemporal suitability maps and probabilistic risk maps for spatial infection spreading. This framework is capable of vector-borne disease risk assessment without historical incidence data and can be a useful tool for preparedness with accurate human movement data.

INTRODUCTION

In modern times, threats posed by infectious diseases have become a significant public health concern due to increasing connectivity.[1] Emerging infectious diseases (e.g., H7N9, H5N1, and Ebola) as well as endemic diseases (e.g., dengue, chikungunya, and measles) pose a severe threat to human health and life.[2] Some infectious diseases have high mortality and morbidity rates (Ebola), and some of these diseases lack treatments or vaccines (dengue).[3,4] Infectious diseases are creating pandemics due to globalization. For example, 2019 had experienced major dengue outbreaks in many countries in the world, including Southeast Asia and Latin America. Other emerging and endemic infectious diseases had also shown to spread rapidly across the globe in recent times. Therefore, accurate risk assessment of disease outbreak is highly important for preparedness in this modern world. Risk has been defined by medical epidemiologists and health organizations[5] as disease development probability within an individual in a specified time interval. Risk is also defined as the potential adverse consequences to human life, health, property, or the environment[6] of unwanted phenomena (disease/event). Accurate risk-assessment models have the potential to improve epidemic prevention and control capabilities.

Because of the complexity in vector-borne disease spreading, risk is often associated with vector or host suitability, the basic reproduction number, vectorial capacity, vector prevalence, or incidence history.[5,7–9] One method to assess risk is by detecting disease outbreaks from surveillance data—a retrospective approach that has a limitation of allowing enough time for preparedness.[10] However, some areas have limited resources for the surveillance system. Especially in developing countries, collected data may not be adequate as most people choose not to use medical facilities unless they have severe conditions. Therefore, disease surveillance data–dependent risk assessment is not always efficient with unreported or underreported incidences. Because of this, researchers have developed other risk-assessment methods to overcome the problem with retrospective methods. The impact of climate change on vector survival, suitability, and pathogen transmission has been assessed for vector-borne diseases in numerous research projects.[11–19] Unfortunately, very limited research has included spatial and temporal heterogeneity of weather conditions, population demography, and movement information in the risk-assessment models.

In this context, we develop a risk-assessment framework incorporating the aforementioned significant elements for vector-borne diseases. In this article, we focus on mosquito-borne diseases to demonstrate the risk-assessment framework as they comprise most of the vector-borne diseases. This article has three major contributions. The first is the formulation of a spatiotemporal, network-based risk-assessment framework by incorporating climate data and demographic information, especially for regions with unreported or underreported incidence data. The second is deriving a spatiotemporal suitability map of competent mosquito species in the disease transmission with only temperature data. The third is the development of spatial risk maps for disease transmission, showing the relative risk of each location compared with others. In addition, identification of the significant-incidence window and peak-incidence period is performed by comparing simulation results with data (when available). A serotype analysis is conducted to identify contributing factors for the year-to-year difference in incidence data. This novel risk-assessment method is capable of incorporating both human movement

*Address correspondence to Lee W. Cohnstaedt, United States Department of Agriculture, Arthropod-borne Animal Diseases Research, 1515 College Ave., Manhattan, KS 66502. E-mail: Lee.Cohnstaedt@usda.gov

and contact patterns, as well as impacts of weather factors in human–mosquito interaction.

Finally, an application of the novel framework is presented for dengue spreading in Bangladesh. A spatiotemporal network is developed for human movement in Bangladesh using demographic information, and 1-month– and 2-month–lagged climate (temperature and rainfall) data. A map for the spatiotemporal suitability of human–mosquito interaction and spatial dengue transmission risk maps are obtained from simulation results. Simulation results match closely with the significant-incidence window and peak-incidence period with Bangladesh dengue transmission dynamics. The year-to-year data variability shows a correlation with the dominant serotype. The combined knowledge obtained from the framework (i.e., significant-incidence window, peak-incidence period, risk map, and spatiotemporal suitability map) provides a guideline to public health personnel in prioritizing spatiotemporal resource allocation to reduce/prevent dengue transmission. Risk maps are developed incorporating generalized human movement data in the spatiotemporal network, and have the adaptability to include actual and accurate movement data.

MATERIALS AND METHODS

Risk-assessment framework. Our novel risk-assessment framework couples a spatiotemporal network-based approach with a compartmental disease model and a spatiotemporal spreading algorithm. The risk-assessment framework has the following five different components:

1. Compartmental model
2. Pathogen transmission model with climate data
3. Spatiotemporal network
4. Spatiotemporal spreading algorithm
5. Risk calculation

Each of these components is described in subsequent parts of this section.

Compartmental model. Compartmental models express transitions of the host population from one disease state/compartment to another.[20] These compartments are, for example, susceptible, exposed, infectious, recovered, removed, vaccinated, and alert. Some parameters govern inter-compartmental transitions. The transition rate from susceptible to exposed/infected compartment (transmission rate) is the most crucial parameter for the vector-borne disease model. The transmission rate is correlated with climate/weather-dependent factors such as vector abundance and host–vector interactions. Therefore, we incorporate climate data in the transmission rate for the risk-assessment framework developed.

Pathogen transmission model with climate data. The transmission rate for vector-borne diseases has a complicated relationship with the environment and the host. For example, mosquito abundance and its interaction with the host population cause the transition from susceptible to infected (or exposed) states. When the mosquito population is the vector for disease, temperature and rainfall data are used to develop the correlation between mosquito abundance and its interaction with the host population. This relationship can be expressed as *vectorial capacity*—a parameter governing the spread of infection from an infected to a susceptible host via vectors. Vectorial capacity is a parameter that governs the spreading of infection from an infected host to a susceptible host via vectors. During the 1960s, Garret-Jones formulated the vectorial capacity, the average number of secondary cases of a disease (e.g., malaria and dengue) arising from each primary infection in a defined population of susceptible hosts.[21,22] As vectorial capacity has been used to transmit the pathogen by Garret-Jones, subsequent literature has also used it for simulation using a network-based model. Scoglio et al.,[23] used vectorial capacity to express the pathogen transmission for Rift Valley fever in cattle using the vectorial capacity for *Culex* mosquitoes. Riad et al.[24] and Sekamatte et al.,[25] have also used vectorial capacity to express pathogen transmission via mosquitoes for Rift valley fever and Japanese encephalitis, respectively. The expression for vectorial capacity is given in Equation (1).

$$V_c = \frac{ma^2 b_h b_m c^{-\mu_m n}}{\mu_m}. \tag{1}$$

Vector parameters used are 1) average daily vector biting rate (a), 2) probability of vector-to-human transmission per bite (b_h), 3) probability of human-to-vector infection per bite (b_m), 4) duration of the extrinsic incubation period (n), 5) vector mortality rate (μ_m), and 6) mosquito vector density with respect to the host (m).[21,22,26] These parameters are specific for the mosquito species and the concerned disease. We choose *Aedes aegypti*, one of the most competent mosquito species in transmitting dengue, Zika, chikungunya, yellow fever, and other severe diseases, to model the transmission rate/vectorial capacity. Except for mosquito vector density with respect to the host, all other parameters in Equation (1) can be calculated empirically using spatiotemporal temperature data.[26] The following empirical formulas are used to calculate temperature-dependent parameters, where T is the temperature in degrees Celsius.

1. Biting rate (a): Liu-Helmersson et al.[26] and Scott et al.[27] developed the following empirical equation from numerous experimental data to model the relationship between temperature and average blood meal frequency of the female *A. aegypti.*

$$a(T) = 0.0043T + 0.0943 \ \left(\text{day}^{-1}\right). \tag{2}$$

2. Probability of vector-to-human transmission per bite (b_h): The empirical equation for the probability of human infection was expressed with the following thermodynamic function.[26,28]

$$b_h\ (T) = 0.001044T(T - 12.286)\sqrt{32.461 - T},\ \text{for}\ 12.286\,^\circ\text{C} \le T \le 32.461\,^\circ\text{C}. \tag{3}$$

3. Probability of human-to-vector infection per bite (b_m): Lambrechts et al.[29] derived the relationships between temperature and the probability of infection based on empirical data for several *A. aegypti*–borne diseases.

$$b_m(T) = \begin{cases} 0.000729T - 0.9037 & (12.4\,^\circ\text{C} \le T \le 26.1\,^\circ\text{C}) \\ 1 & (26.1\,^\circ\text{C} \le T \le 32.5\,^\circ\text{C}). \end{cases} \tag{4}$$

4. Duration of the extrinsic incubation period (n): An exponential function was used to fit experimental data for the extrinsic incubation period.[29,30]

$$n(T) = 4 + e^{5.15 - 0.123T}. \quad (5)$$

5. Vector mortality rate (μ_m): Yang et al.[31] developed a fourth-order polynomial equation to fit experimental data to model the mortality rate in relation to temparature.

$$\mu_m(T) = 0.8692 - 0.1590T + 0.01116T^2 - 3.408 * 10^{-4}\ T^3 + 3.809 * 10^{-6}T^4. \quad (6)$$

6. Mosquito vector density with respect to the host (m): The mosquito vector density with respect to the host mostly depends on the rainfall. Therefore, in this article, this parameter is expressed as proportional to the weekly average rainfall.[32] We have normalized the average weekly rainfall in each location before using it in the vectorial capacity model.

Spatiotemporal network. To account for the spatiotemporal heterogeneity of the disease transmission risk with changing weather conditions, we propose a spatiotemporal network. The spatiotemporal network is developed using host demographic information such as population density, distribution, and movement. In this network, nodes represent individuals within spatially homogeneous locations, and links represent movements within and between these locations. The network is spatially explicit and has multiple temporal realizations to represent heterogeneities in weather conditions with time and space. This network is then combined with a spatiotemporal spreading algorithm to simulate the spatiotemporal transmission of the infection. The network is periodically updated to reflect changing weather conditions in the spatiotemporal spreading algorithm.

Spatiotemporal spreading algorithm. Nodes influence each other through statistically independent pairwise interactions in most network-based models. Sahneh et al.[33,34] developed the *generalized epidemic modeling framework* (GEMF) for stochastic spreading processes over complex networks based on these independent pairwise interactions. The GEMFsim tool was later developed for numerical simulation of GEMF-based models by implementing the Gillespie algorithm.[33] The combined state of all nodes in a network can be described as a random variable $X_N(t) = x_1(t), x_2(t), \ldots, x_i(t)$, where $x_i(t)$ is the state (compartment) of node i at time t. Transition time from one state to another is expressed as an exponential distribution with a transition rate $\sigma_n(x_n \to J)$, where J is the destination state after the transition. This transition can be node based (dependent only on the node state $x_i[t]$) or edge based (dependent on the combined network state $X_N[t]$). After a transition occurs, the combined network state will change, and therefore, edge-based transition rates will change. However, node-based transition rates remain constant. *Generalized epidemic modeling framework*sim accounts for changes in transition rates due to the change in the combined network state. However, the GEMFsim does not account for the temporal variation of the transition rates due to external factors (weather conditions or human activities).

The temporal variability of the transition rate is crucial for simulating vector-borne disease transmission. Therefore, it is required to adapt the Gillespie algorithm in the GEMFsim to account for the changing rates. In this article, we incorporate the nonhomogeneous Gillespie algorithm in the GEMFsim, which works for exponential event distributions and nonconstant transition rates.[4,35,36] The modified spreading algorithm is capable of periodically changing transition rates to reflect the temporal heterogeneity of vector-borne disease transmission.

Risk calculation. As the spreading process in our spatiotemporal network is highly stochastic, we need to perform an adequate number of simulations. We keep track of each node's status and count the numbers of simulations in which a particular node is infected. This count is later used to calculate the risk of the spatial disease spreading. The formula for risk calculation is $R_j = \frac{\sum_{n=1}^{N_j} I_n}{N_j}$, where R_j = spreading risk in location j, I_n = number of simulations where node n is infected, $N_{\text{simulation}}$ = number of simulations, and N_j = total number of population in jth location. The calculated risk is normalized for comparison with the risk of different spatial locations.

Application of the risk-assessment framework for Bangladesh dengue incidence. Dengue is transmitted by *Aedes* mosquitoes in tropical and subtropical regions and may cause a wide range of manifestations from asymptomatic infections to deaths.[37] Arbovirus transmission is known to be driven by the interplay of gender, age, and travel of individuals. In addition, the transmission also depends on the type of host community (urban/rural), mosquito abundance, and use of mosquito control measures.[38–41] Understanding the relative importance of these factors is required for accurately assessing the risk of dengue.

A recent dengue outbreak in Bangladesh with a record number of cases drew close attention to assess the transmission risk. Bangladesh has a history of dengue incidence dating back to 1960, and a major outbreak occurred in 2000.[42–44] Since 2000, the Ministry of Health and Family Welfare of the People's Republic of Bangladesh has been recording clinical cases, which are reported annually. Recent urbanization throughout the tropical world has accelerated dengue spreading as *A. aegypti*—the primary vector for dengue transmission—lives in densely populated human-made environments.[45] Bangladesh is a densely populated country with rapid urbanization, which provides a conducive environment for mosquito populations. Therefore, a risk-assessment tool for dengue transmission has become very important in Bangladesh.

Several studies have used historical time series data for reported dengue incidence. However, getting accurate data on dengue infection is surprisingly difficult. Only 11–32% of infected people are likely to have symptoms, with just a few being sick enough to require formal medical treatment.[45,46] Misdiagnosis and underreporting are common for cases requiring medical care as well. Therefore, risk assessments based on clinical case counts are not always useful, and may just reflect differences in access to healthcare, diagnostics, and the ability to report cases.[47] Record-keeping also requires significant resources not always available in Bangladesh. Therefore, our developed framework, which does not require incidence data, can be a useful tool for assessing spatial dengue transmission risk and spatiotemporal suitability in Bangladesh. The adapted framework for Bangladesh dengue spreading is presented in the following text.

Compartmental model for dengue. When the dengue virus (DENV) enters the bloodstream of a susceptible person via infected mosquito bites, the individual becomes exposed to the disease. After a specific time for viral replication, the

exposed individual becomes infectious. The infectious individual finally transitions to the removed state after recovery or death. Therefore, four specific phases/states concern the disease—susceptible, exposed, infected, and removed. The model is called *SEIR*. Its inter-compartmental transitions are independent Poisson processes with transitions rates expressed in Equations (7)–(9).

$$Pr(x_i[t+\Delta t]=E|x_i[t]=S,\ X_N[t])=\beta_i(t)Y_i\Delta t+o(\Delta t), \quad (7)$$

$$Pr(x_i[t+\Delta t]=I|x_i[t]=E,\ X_N[t])=\delta\Delta t+o(\Delta t), \quad (8)$$

$$Pr(x_i[t+\Delta t]=R|x_i[t]=I,\ X_N[t])=\gamma\Delta t+o(\Delta t). \quad (9)$$

In these equations, $x_i(t + \Delta t) = S$, E, I, and R express the probability of node i occupying the susceptible, exposed, infected, or removed state at time $(t + \Delta t)$, respectively. $X_N(t)$ is the combined network state at time t. The transition rate from susceptible to exposed state is an edge-based transition, which is also time variant because of its dependency on weather conditions. We express this time-variant parameter $\beta_i(t)$ with vectorial capacity for node i at time t, which is calculated from spatiotemporal weather conditions. Y_i is the set of infected neighbors of node i within the spatiotemporal network at time t. Parameter δ is the intrinsic incubation rate, which governs the transition from exposed to infected state. The transition from infected to removed state is expressed with the removal rate γ. Incubation rate δ and removal rate γ are node-based transition rates whose values are assumed equal to 0.17 and 0.14, respectively, and are time invariant in this work.[48,49] These values of δ and γ reflect the means of exponentially distributed parameter values used in the spatiotemporal spreading process.

Pathogen transmission model with Bangladesh climate data. Transmission rate β is modeled with weekly average temperature and rainfall data in Bangladesh. Climate data for Bangladesh are collected from CLIMATE-DATA.ORG.[50] The upazila-level spatial unit is used in this work for network development. Climate data are used to calculate each parameter in the vectorial capacity in Equation (1). All parameters, except for mosquito vector density with respect to the host, are calculated from the weekly temperature data. The mosquito vector density parameter is assumed proportional to the weekly average rainfall, and a proportional constant is assumed to reflect a realistic outbreak scenario in Bangladesh.

An urbanization factor is assumed to reflect the suitability of an *Aedes* mosquito habitat for dengue transmission. The population density is used to classify the urbanization level of each location. Three urbanization factors are used to reflect back-country (population density < 1,000/km^2), rural (1,000 ≤ population density < 3,000/km^2), and urban (population density ≤ 3,000/km^2) locations. The final transmission rate used in simulating dengue transmission is equal to vectorial capacity multiplied by the urbanization factor.

Spatiotemporal network for Bangladesh. Developing a network is a crucial part of the risk assessment framework. The spatial structure in Bangladesh is as follows:

1. Administrative level 1, eight divisions.
2. Administrative level 2, 64 districts.
3. Administrative level 3, 544 upazila.

For network development, we have used the administrative level 3 spatial resolution. Therefore, Bangladesh is divided into 544 spatial locations before creating the network. Population data are collected for each spatial location from the city population website.[51] The population in each location is scaled by 10,000 to reduce the computational burden during simulations.

We assume an Erdos–Renyi network within each upazila, where links are created with a probability of 0.2. Inter-upazila links are created using an exponential dispersion kernel. We use the kernel function e^{-kD} for link generation, where k is a constant, and D is the distance between the source and destination location. We choose the value of k = 0.1 for creating the spatiotemporal network. District-level and division-level human movement, along with the exponential dispersion, are incorporated to reflect human movement patterns. District-level human movement is incorporated by generating links between the capital city Dhaka and all-district cities. Links are created between Dhaka and all-division cities to include division-level human movement in the network. Figure 1 demonstrates a simplified outline of the network.

Bangladesh is a small country, having only 147,500 km^2 of area. However, there are still spatial and temporal difference



FIGURE 1. A simplified network diagram for Bangladesh. Each black circle represents the network within an upazila, whereas lines between circles express human movement. Circle sizes are scaled according to human movement (node degree) for that location. Greater circle size indicates a greater amount of human movement flow.

between temperature and rainfall throughout the country. This spatiotemporal weather variation is very important for accurately representing dengue transmission. We included a spatiotemporal heterogeneity in the created network due to weather patterns, that is, pathogen transmission rates on each link. The literature shows the correlation between a 1-month– and 2-month– lagged temperature, rainfall, and dengue occurrence in Bangladesh.[52] The heterogeneity in the weather patterns is reflected on a weekly value of transmission rate (β) calculated using both 1-month– and 2-month–lagged temperature and rainfall data.

We create two instances of the spatiotemporal network for two different outbreak types: first, major outbreaks spreading throughout the whole country, and second, minor outbreaks spreading only within major divisional cities. We explicitly incorporate district-level human movement for major outbreak scenarios with the exponential dispersion kernel. Division-level human movement is incorporated for minor outbreak scenarios along with the exponential dispersion. On developing the spatiotemporal network, we apply the stochastic spreading algorithm for risk assessment.

RESULTS AND DISCUSSION

Assessed risk in this work is a combination of weather-dependent spatiotemporal suitability and risk maps from the network-based model. Therefore, we present our results in the following two sections as

1. spatiotemporal suitability of dengue transmission in Bangladesh
2. risk maps for dengue transmission in Bangladesh

Spatiotemporal suitability of dengue transmission in Bangladesh. Spatiotemporal suitability expresses spatial and temporal suitability of the vector's (mosquito's) survival and function in pathogen transmission. When comparing dengue epidemic potential over time and space, it is preferable to use the relative vectorial capacity,[26] expressed as the vectorial capacity relative to the vector-to-human population ratio, and formulated as $rV_c = \frac{a^2 b_h b_m c^{-\mu_m n}}{\mu_m}$.[26] All parameters are temperature dependent in the relative vectorial capacity definition. A higher relative vectorial capacity indicates a higher potential for a dengue epidemic. We calculate the relative vectorial capacity for the *Aedes* mosquito in Bangladesh to infer the suitability of dengue spreading. A threshold value for relative vectorial capacity exists beyond which *Aedes* mosquitoes can function properly in transmitting dengue infection. A relative vectorial capacity greater than 0.61 indicates the suitability of dengue spreading.[26] The spatiotemporal suitability maps for dengue spreading in Bangladesh are presented in Figure 2 for each month of the year. The relative vectorial capacity is found to be higher than the threshold value (0.61) for many months of the year in almost all locations (red and yellow regions).

It is evident from Figure 2 that months between January and March are poorly suited for dengue transmission. In April, some southern parts of the country, as well as the capital city of Dhaka, become highly suitable. Other parts of the country, except the northern most corner, become moderately suitable. Starting from May, the whole country becomes highly suitable, and the situation remains similar until November. The most northern part of the country becomes poorly suitable, whereas the capital and southern part stay highly suitable in December. After December, the whole country becomes poorly suited again, which continues until April.

Dengue incidence data since 2000 show cases throughout the year, which supports our results for the suitability of dengue transmission. Bangladesh is a tropical country, which provides a suitable temperature for mosquito survival year-round. However, for temperate regions with widely varying temperatures, mosquitoes may not be able to survive in the coldest months.

Risk maps for dengue transmission in Bangladesh. The network-based model enables us to simulate the spreading process within the spatiotemporal network described earlier. Every year, the dengue outbreak in Bangladesh shows a different trend. Some years cases are reported from most parts of the country, whereas some years outbreaks are mostly limited within divisional cities. Depending on the spreading, outbreaks are divided into two categories—major outbreaks, with widespread dengue cases, and minor outbreaks, with cases mostly in some divisional cities. Therefore, two distinct simulation scenarios are assumed to match the two different outbreak types. Simulations are performed with parameters calculated using both 1-month– and 2-month–lagged temperature and rainfall data, which are presented in the subsequent parts of this section.

Scenario 1: Major outbreak. A major outbreak is defined when dengue cases are widespread throughout the whole country. For major outbreaks, the network is generated with district-level human movement incorporated. Dhaka is the capital of the country; therefore, frequent movement among all district towns is assumed to and from Dhaka. We performed 1,000 iterations for each scenario to account for stochasticity in the simulation results. Simulations are started with an entirely susceptible population with one infected human in the initial outbreak location. The initial outbreak location is also changed to see the impact of human movement on transmission risk. Simulation results for 1-month–lagged climate data, and two initial conditions—Dhaka and Chittagong outbreaks starting—are presented in Figure 3.

Figure 3A and D show the fractions of infected people and cumulative infected people with a 95% CI when the simulation started in Dhaka and Chittagong, respectively. The curve for the average number of infected shows two peaks in Figure 3A and D. The first peak refers to the rapid spreading within the vicinity of the initial outbreak location, and the second peak represents the widespread outbreak. For novel vector-borne diseases, the infections start with a single infected human or a single infected mosquito. If we assume a single infected human started the infection, then the infection will start locally with competent mosquitoes biting the infected person. As mosquitoes have a short flying range, the infection will be local at the beginning, and this is evident from the smaller initial peak. Size of the peak depends on population size, level of urbanization, and area of the outbreak location. As Dhaka is densely populated, an outbreak starting there will result in a pronounced initial peak due to only local transmission (Figure 3A). We started our simulation in April as it is the first month of high suitability for the *Aedes* mosquito within the year for Bangladesh. However, the suitability increases with time, as shown in Figure 2, and the infection starts spreading to distant locations because of human movement as well as higher suitability. Hence, we have another peak in August with

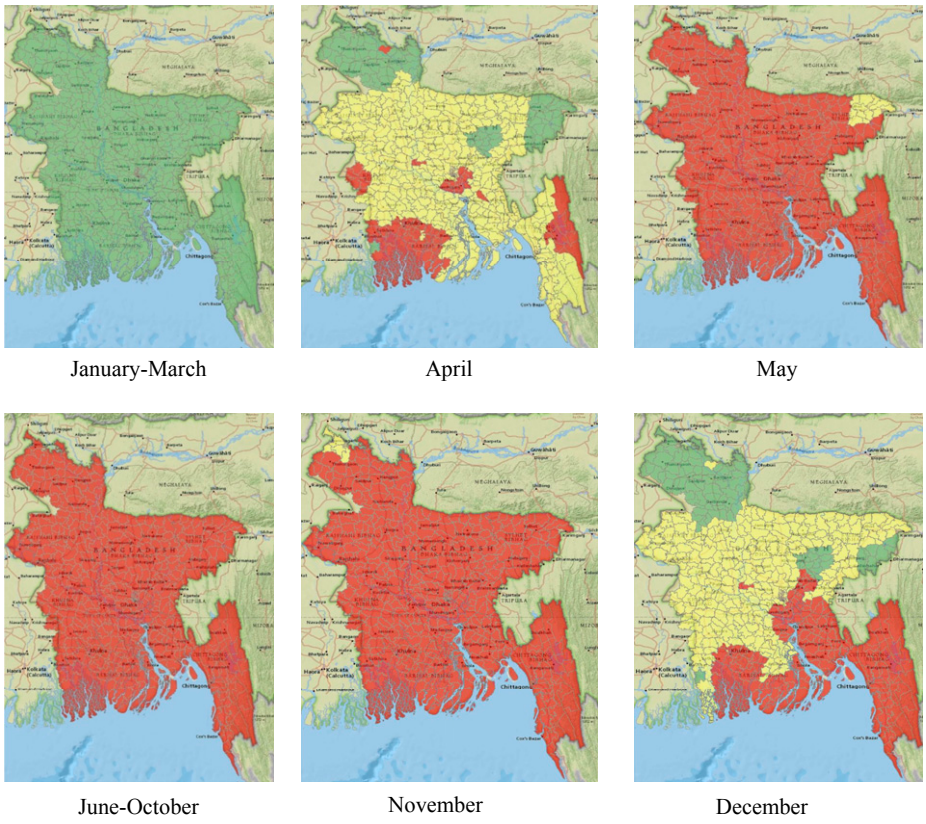


FIGURE 2. Spatiotemporal suitability maps for dengue transmission in Bangladesh based on temperature. The map represents suitability in the following manner: green (lowly suitable), yellow (moderately suitable), and red (highly suitable). This figure appears in color at www.ajtmh.org.

numerous cases countrywide. Although Figure 2 shows all locations are highly suitable from May to November, the underlying mosquito-dependent transmission rate keeps increasing until August and attains higher values in June–August. Therefore, despite human movement from Dhaka to other locations all year, because of the lower value of transmission rates, no widespread infections start until June. With the higher value of transmission rates, the number of infected individuals keeps increasing from June and attains its peak in August. Around 16% of the total population in the network becomes infected in this outbreak scenario. Almost 70% of the total infections are observed within the period of July–September. The peak consists of 2.5% of the entire population in our developed network. These infected people during August may require hospital care, which will be a huge burden on the healthcare system. Therefore, proper measures should be taken, and more resources should be allocated for dengue health care during July–September. Figure 3D shows the dynamics when dengue infection started in Chittagong. Total fractions of cumulative infected people and infected people during peak times are both significantly lower for this outbreak scenario, which can be attributed to the fact that Chittagong is not very densely populated as well as not well connected to the whole country as is Dhaka. Therefore, both the initial smaller peak and the second higher peak are smaller than the peaks in the Dhaka outbreak scenario. When the infection starts in Chittagong, around 6% of the population becomes infected compared with 16% in the Dhaka outbreak scenario. Therefore, starting location of the infection is a crucial determining factor of the extent and dynamics of the epidemic.

Figure 3B and E show the histogram with fractions of cumulative infected humans in the Bangladesh dengue network and the number of simulations performed. The *x*-axis represents fractions of infected humans in the simulation, and the *y*-axis shows the number of simulations where a particular infection size is obtained. Figure 3B shows that almost 80% of simulation results are in the 10–20% of infected humans in the representative network when the initial outbreak happens in Dhaka. This accounts for the narrower CI in Figure 3A and E expresses around a 20% probability of 10–20% of humans being infected. This variability in the fractions of infected individuals accounts for the wider CI in the simulation results in Figure 3D.

Transmission dynamics presented in Figure 3 are obtained using the network with district-level human movement incorporated, and 1-month–lagged temperature and rainfall data. However, when we perform simulations with 2-month–lagged temperature and rainfall data, the simulation is started in May, and the peaks are delayed by a month. The analysis of Bangladesh dengue incidence data since 2000 showed the occurrences of peaks in July–October.[42,43] Therefore, 2-month–lagged data also show significant similarity with the actual incidence data. Risk maps are similar for both 1-month–lagged and 2-month–lagged data. Simulation results for 2-month–lagged data are presented in the supporting materials.

Scenario 2: Minor outbreak. Dengue infections are often confined within divisional cities, and a widespread outbreak

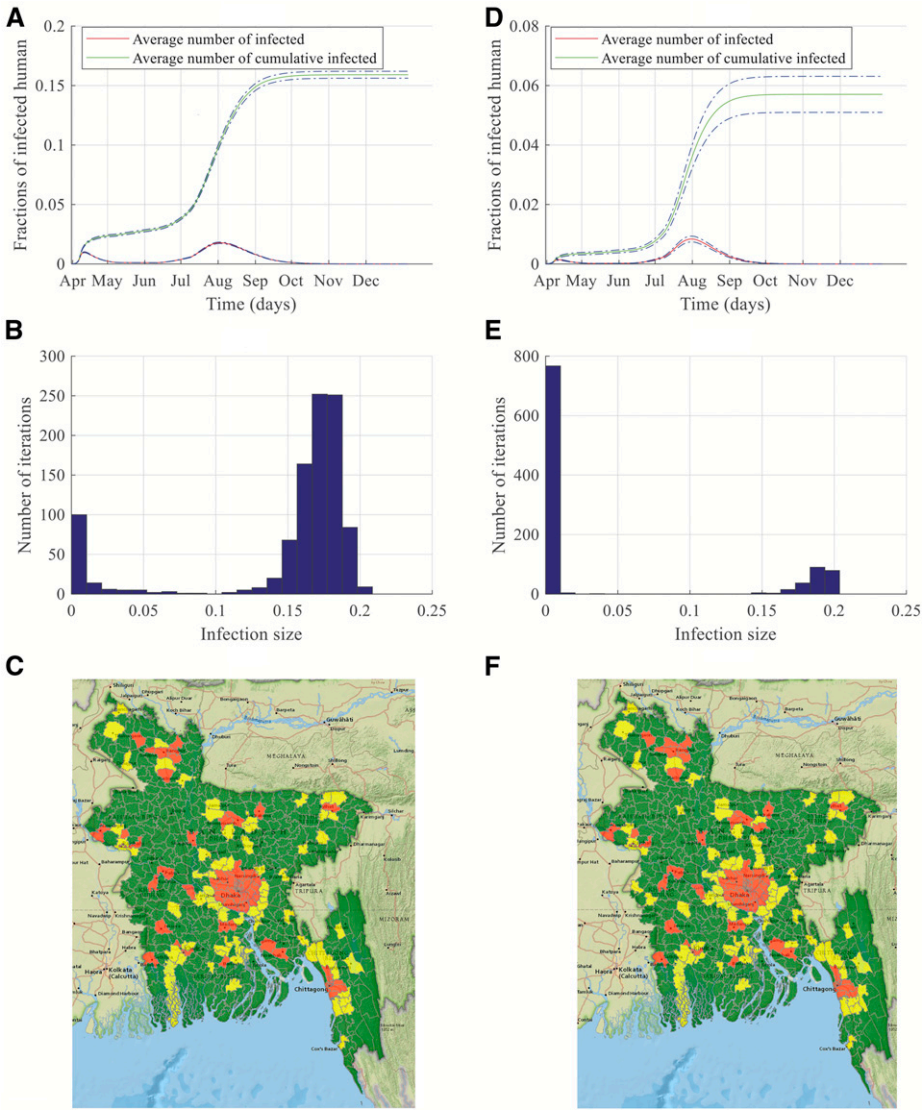


FIGURE 3. Simulation results and risk maps for dengue transmission in Bangladesh for a major outbreak. Left-side panels are results of simulations started in Dhaka, whereas right-side panels are results of simulations started in Chittagong. (**A** and **D**) show dengue transmission dynamics, (**B** and **E**) present histograms of the number of simulations and infection size, and finally, (**C** and **F**) display risk maps for dengue infection. This figure appears in color at www.ajtmh.org.

does not happen. Therefore, we propose another scenario where we explicitly incorporate human movement from the city of Dhaka to other divisional cities, along with the distance-based exponential movement kernel. We call this scenario a minor outbreak as the infection does not spread countrywide. We perform simulations for both 1-month– and 2-month–lagged climate data.

The results are presented in Figure 4, when simulations are performed with 1-month–lagged climate data and the same initial conditions as major outbreak scenarios, namely, starting simulations in Dhaka and Chittagong.

Figure 4A and D show a similar trend as shown in Figure 3A and D. For 1-month–lagged data, a major peak is observed in August with rapidly increasing infections from July–September. However, comparing Figures 3 and 4, we can see that both the value of the fraction of infected people during the peak infection time and the whole outbreak period are smaller during minor outbreaks.

Human movement can also be considered as a crucial factor in vector-borne disease spreading, although the pathogen transmission does not happen via direct physical contact. An exposed/infected person may move/travel to a different location and become infectious after reaching the destination. That person can be bitten by a competent local mosquito and may start a local outbreak in the destination location. Therefore, mosquitoes are responsible for local transmission, whereas human movement is mostly responsible for long-distance pathogen transmission during a period of higher suitability. The reduction in the number of infections in the minor outbreak scenario can be attributed to the reduced human movement volume within the network rather than with the major outbreak scenario.

It is evident from the comparison between panels (a) and (d) of Figures 3 and 4 that the major dengue-spreading period in Bangladesh is June–September. Identifying this time window for significant transmission is crucial for public health officials.

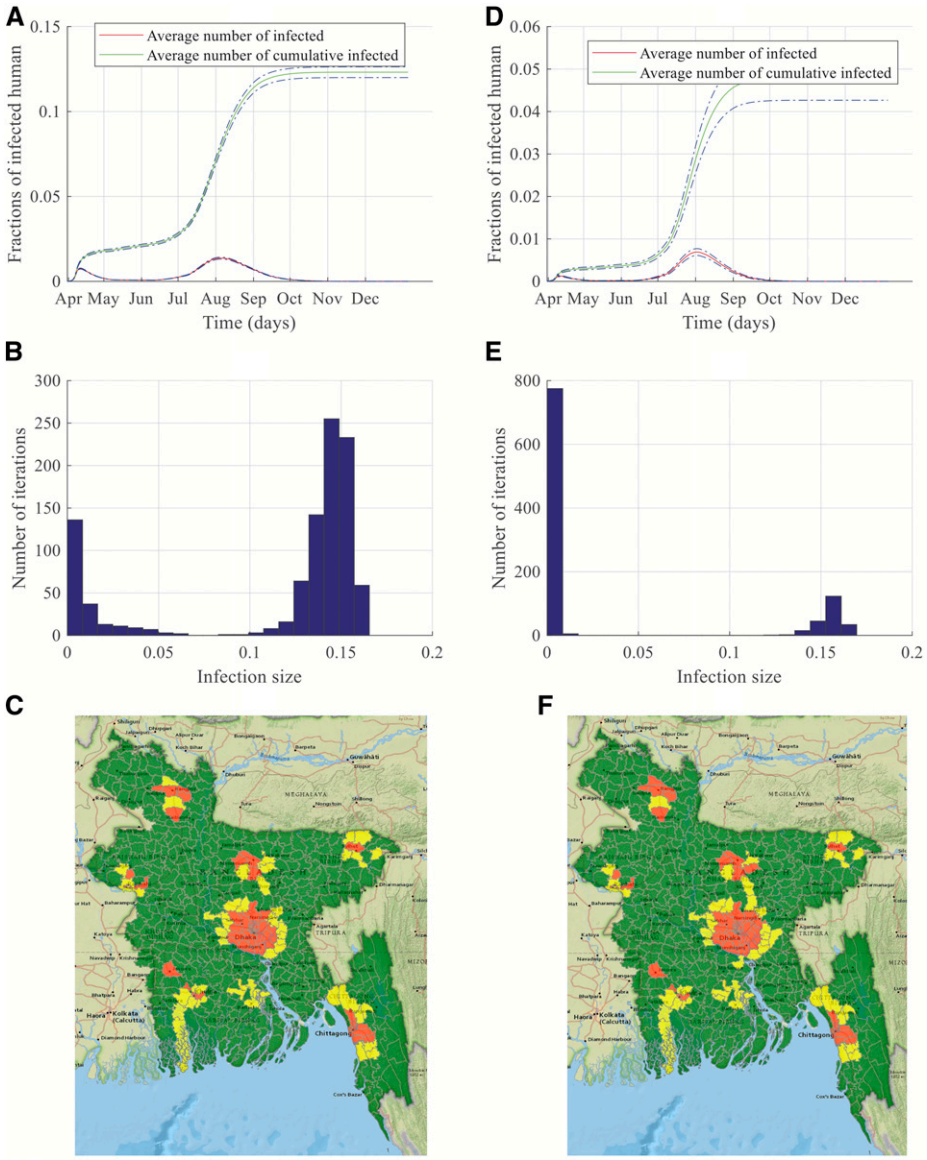


FIGURE 4. Simulation results and risk maps for dengue transmission in Bangladesh for a minor outbreak. Left-side panels are results of simulations started in Dhaka, whereas right-side panels are results of simulations started in Chittagong. (**A** and **D**) show dengue transmission dynamics, (**B** and **E**) present histograms of the number of simulations and infection size, and finally, (**C** and **F**) display risk maps for dengue infection. This figure appears in color at www.ajtmh.org.

Identification of the significant-incidence window will enable them to take prompt action during the surge of infection. The major action for controlling a vector-borne disease is always controlling the vectors. Therefore, control measures need to focus mostly on the significant-incidence months to reduce the mosquito population when resources are inadequate for the whole year.

Figure 4C and F show the normalized risk maps for dengue spreading during a minor outbreak. It is evident from the figures that high-risk areas are confined within major-division cities and their nearby locations in contrast to Figure 3C and F, where many locations throughout the country are at high risk. This reduction in the spreading risk can be attributed to reduced human movement in the minor outbreak scenario. Therefore, making people aware of human movement's impact on long-distance travel through social media or radio/TV broadcasting would help contain the epidemic. Simulations are also performed for 2-month–lagged climate data, and all results are like those obtained with 1-month–lagged data. These simulation results are presented in the supporting materials.

Our proposed framework is a generalized risk-assessment tool based on climate and demographic data, which can be used for risk assessment, especially in regions with unreported or underreported incidence data. Risk maps developed in this work are generated with the generalized concept of human movement within Bangladesh. The framework can incorporate more detailed and accurate human movement data. The incorporation of detailed movement data will provide a more accurate assessment of the transmission risk at each location. Once proper and accurate movement data are incorporated into the network, control measures

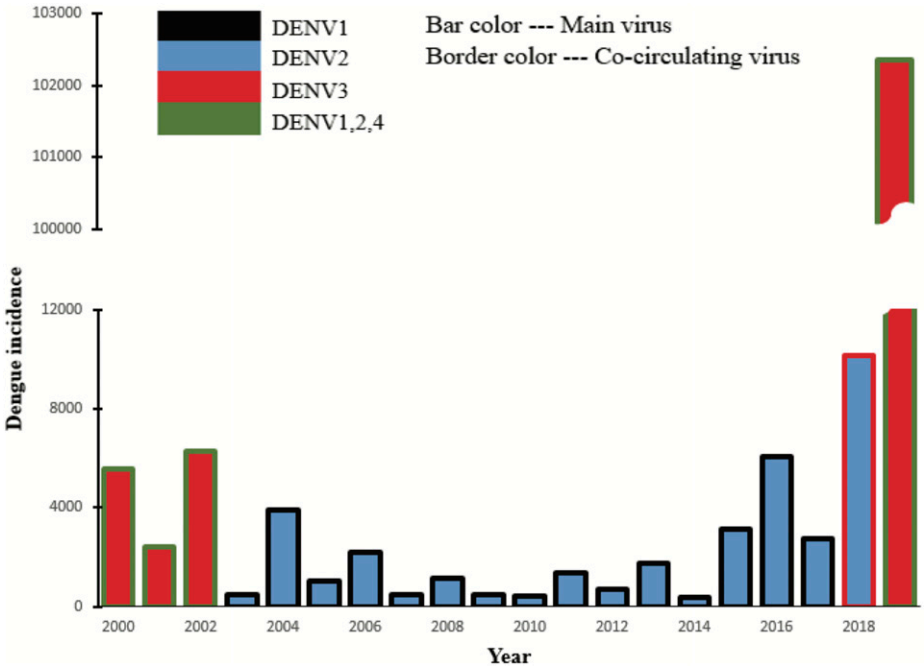


FIGURE 5. Serotype analysis of dengue spreading in Bangladesh since 2000. The bar chart presents the number of dengue cases with circulating serotypes each year in Bangladesh. The main bar color represents the dominant serotype, whereas the border represents other circulating serotypes. This figure appears in color at www.ajtmh.org.

should be applied to high-risk areas first, followed by medium- and low-risk areas, depending on availability of resources.

Serotype analysis. Since 2000, dengue cases arise every year in Bangladesh. However, the number of cases varies from year to year. Importantly, the available data concern only hospitalized and reported cases of dengue. In this section, we show the existence of a correlation between the number of cases and the circulating DENV serotypes. Dengue fever can be caused by any of four genetically related DENV serotypes (DENV1, DENV2, DENV3, and DENV4).[53] After recovering from infection with one dengue serotype, a person has immunity against that particular serotype.[53] Unfortunately, the person can be infected again with any of the remaining three dengue serotypes.[54] Subsequent infections often put individuals at greater risk for severe dengue illnesses than those who have not been previously infected.[55] A bar graph of yearly dengue incidences in Bangladesh with serotypes is presented in Figure 5.

The primary bar color represents the dominant circulating serotype, whereas the border color represents co-circulating serotypes. The *x*-axis shows the year, and the *y*-axis shows the number of reported dengue cases in Bangladesh. The bar and border colors are as follows: black for DENV1, blue for DENV2, red for DENV3, and green for DENV1, 2, and 4 together. The height of each bar corresponds to the number of cases in the corresponding year. Figure 5 shows the dominant circulating serotype was DENV3, with the other three co-circulating during the years 2000–2002. Within the period 2003–2016, DENV2 was the dominant serotype and DENV1 was co-circulating.[54] The DENV3 reemerging in 2017 resulted

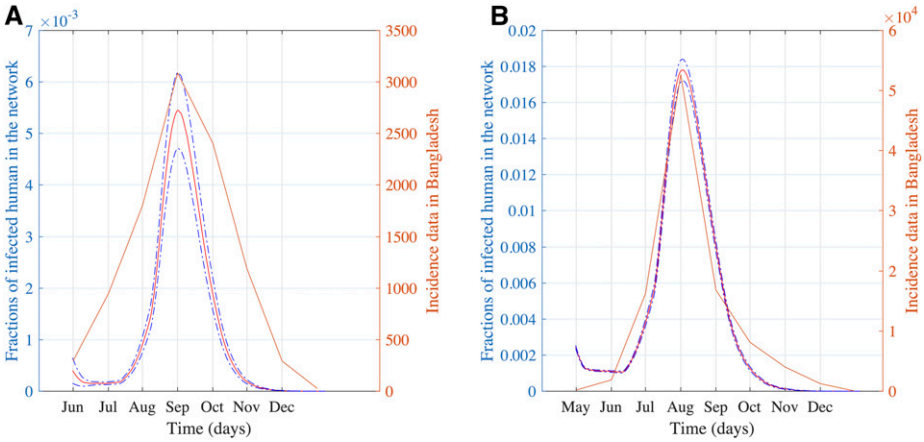


FIGURE 6. (**A**) Comparison of peak time from our simulation with incidence data during a minor outbreak in 2018 and (**B**) comparison of peak time from our simulation with incidence data during a major outbreak in 2019. This figure appears in color at www.ajtmh.org.

in a significant increase in the number of cases during 2018. However, DENV2 was still the dominant circulating serotype in 2018.[54] If a serotype is circulating long enough within a population, all recovered people may become immune to that specific serotype. An increase in the number of cases in 2018 can be attributed to the reemergence of DENV3. DENV3 became the dominant circulating serotype in 2019, which created an extreme and unprecedented surge in the number of infections. More than 100,000 cases were reported in 2019—which is more than double the number of combined cases in the previous 19 years—because of a vast susceptible population for DENV3 serotype. Currently, DENV3 is already in circulation in Bangladesh. Neighboring countries of Bangladesh have all four serotypes. Therefore, Bangladesh is always at risk for all four serotypes, including DENV4. Healthcare personnel should be vigilant to identify dengue patients before the significant-incidence period to identify the circulating serotypes. The introduction of a new serotype will produce a high probability of a surge in dengue infections.

Peak timing validation. The yearly reported cases varied widely for dengue incidence in Bangladesh. For example, there were 10,148 cases in 2018, and more than 100,000 cases were reported in 2019.[54,56] Therefore, a high variability in year-to-year dengue cases exists. From the serotype analysis in the previous section, the unprecedented increase in 2019 can be attributed to the DENV3 circulation. The yearly number of dengue cases is a complex combination of circulating serotypes, movement patterns of the human population, and measures taken for mosquito control. For widely varying year-to-year case numbers, we used the peak-incidence time and transmission dynamics to compare our simulation results with the incidence data. Peak incidence happened mostly in August and September in Bangladesh.[54] Simulation results with 1-month–lagged climate data show incidence dynamics with the major peak in August. Two-month–lagged climate data resulted in a peak in September. We compare our simulation with actual incidence data for 2018 and 2019 in Figure 6.

In 2018, the peak of incidence data was observed in September, as shown in Figure 6A. Simulation results for 2-month–lagged climate data show a peak in September in accordance with 2018 incidence data in Figure 6A. Disease dynamics from simulations with 1-month–lagged climate data shows a similar trend with a coinciding peak for 2019 (Figure 6B).

CONCLUSION

Climate data can be used to develop a generalized risk-assessment framework for vector-borne diseases together with demographic data—spatial distribution of individuals and their movement patterns. Therefore, we develop a novel risk-assessment framework with a spatiotemporal network model and a nonhomogeneous Gillespie algorithm, using both climate and demographic data. The assessed risk from this framework comprised spatiotemporal suitability maps and spatial risk maps. Spatiotemporal suitability maps show spatial and temporal suitability of vector-borne transmission. Spatial risk maps represent the disease transmission risk of each location compared with other locations on the map. This framework also identifies high-risk or elevated-risk months, as well as the peak-incidence period within a year.

On development, the framework is applied to the study of dengue transmission in Bangladesh for major and minor outbreak scenarios. The difference between major and minor outbreaks is defined by different levels of human movement to demonstrate the critical role of human dispersal on widespread pathogen transmission. Reduced human migration throughout the country will reduce the spread of infection to divisional cities. We generate the spatiotemporal suitability map and the risk maps for Bangladesh dengue transmission. Simulation results also showed a similar significant-incidence window as well as the peak-incidence period with reported dengue-incidence data in Bangladesh. Serotype analysis indicates the importance of identifying circulating DENV serotypes before the significant-incidence window. The possibility of a major outbreak is associated with the introduction and reemergence of new DENV serotypes. The proposed risk-assessment tool provides guidelines for public health officials to prioritize resource allocation and control measure application according to the estimated risk.

The network-based model has the flexibility to incorporate very detailed mitigation measures in the model. These mitigation measures can be social distancing, reduced human movement, or other disease-specific preventive measures. We will extend this work to explore different mitigation strategies in our future work.

Received May 8, 2020. Accepted for publication November 14, 2020.

Published online January 18, 2021.

Note: Supplemental material appears at www.ajtmh.org.

Financial support: The project was sponsored by the Department of the Army, U.S. Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Ft Detrick (DWFP Grant W911QY-19-1-0004), and NSF/NIH/USDA/BBSRC Ecology and Evolution of Infectious Diseases (EEID) Program through USDA-NIFA Award 2015-67013-23818.

Authors' addresses: Mahbubul H. Riad and Caterina M. Scoglio, Kansas State University, Manhattan, Kansas, E-mails: mahbubriad@ksu.edu and caterina@ksu.edu. Lee W. Cohnstaedt, United States Department of Agriculture, Manhattan, KS, E-mail: lee.cohnstaedt@ars.usda.gov.

This is an open-access article distributed under the terms of the Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

REFERENCES

1. Singh SK, 2019. Infectious diseases, national security and globalisation. *World Aff J Int Issues 23:* 10–23.
2. Rauch S, Jasny E, Schmidt KE, Petsch B, 2018. New vaccine technologies to combat outbreak situations. *Front Immunol 9:* 1963.
3. Hayden FG, Friede M, Bausch DG, 2017. Experimental therapies for Ebola virus disease: what have we learned? *J Infect Dis 215:* 167–170.
4. Ferdousi T, Cohnstaedt LW, McVey D, Scoglio CM, 2019. Understanding the survival of Zika virus in a vector interconnected sexual contact network. *Sci Rep 9:* 7253.
5. Sedda L, Morley D, Braks M, De Simone L, Benz D, Rogers D, 2014. Risk assessment of vector-borne diseases for public health governance. *Public Health 128:* 1049–1058.
6. Thompson KM, Deisler PF, Jr., Schwing RC, 2005. Interdisciplinary vision: the first 25 years of the society for risk analysis (SRA), 1980–2005. *Risk Anal Int J 25:* 1333–1386.
7. Brownstein JS, Rosen H, Purdy D, Miller JR, Merlino M, Mostashari F, Fish D, 2002. Spatial analysis of west Nile virus:

rapid risk assessment of an introduced vector-borne zoonosis. *Vector Borne Zoonotic Dis 2:* 157–164.
8. Romi R, Boccolini D, Menegon M, Rezza G, 2012. Probable autochthonous introduced malaria cases in Italy in 2009–2011 and the risk of local vector-borne transmission. *Eurosurveillance 17:* 20325.
9. Robin C, Lo Iacono G, Gubbins S, Wood J, Newton J, 2013. The accuracy of the national equine database in relation to vector-borne disease risk modelling of horses in Great Britain. *Equine Vet J 45:* 302–308.
10. Gui X, Kou Y, Pine K, Ladaw E, Kim H, Suzuki-Gill E, Chen Y, 2018. Multidimensional risk communication: public discourse on risks during an emerging epidemic. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal, Canada: ACM, 214.
11. Medlock JM, Leach SA, 2015. Effect of climate change on vector-borne disease risk in the UK. *Lancet Infect Dis 15:* 721–730.
12. Medlock JM, Avenell D, Barrass I, Leach S, 2006. Analysis of the potential for survival and seasonal activity of *Aedes albopictus* (Diptera: Culicidae) in the United Kingdom. *J Vector Ecol 31:* 292–305.
13. Caminade C, Kovats S, Rocklov J, Tompkins AM, Morse AP, Colón-González FJ, Stenlund H, Martens P, Lloyd SJ, 2014. Impact of climate change on global malaria distribution. *Proc Natl Acad Sci U S A 111:* 3286–3291.
14. Fischer D, Thomas SM, Niemitz F, Reineking B, Beierkuhnlein C, 2011. Projection of climatic suitability for *Aedes albopictus* Skuse (Culicidae) in Europe under climate change conditions. *Glob Planet Change 78:* 54–64.
15. Fischer D, Thomas SM, Suk JE, Sudre B, Hess A, Tjaden NB, Beierkuhnlein C, Semenza JC, 2013. Climate change effects on chikungunya transmission in Europe: geospatial analysis of vector's climatic suitability and virus' temperature requirements. *Int J Health Geogr 12:* 51.
16. Service M, 1977. Ecological and biological studies on *Aedes* cantans (Meig.) (Diptera: Culicidae) in southern England. *J Appl Ecol 14:* 159–196.
17. Rogers D, Randolph S, 2006. Climate change and vector-borne diseases. *Adv Parasitol 62:* 345–381.
18. Dhiman RC, Pahwa S, Dhillon G, Dash AP, 2010. Climate change and threat of vector-borne diseases in India: are we prepared? *Parasitol Res 106:* 763–773.
19. Gale P, Brouwer A, Ramnial V, Kelly L, Kosmider R, Fooks A, Snary E, 2010. Assessing the impact of climate change on vector-borne viruses in the EU through the elicitation of expert opinion. *Epidemiol Infect 138:* 214–225.
20. Keeling MJ, Eames KT, 2005. Networks and epidemic models. *J R Soc Interf 2:* 295–307.
21. Garrett-Jones C, 1964. The human blood index of malaria vectors in relation to epidemiological assessment. *Bull World Health Organ 30:* 241–261.
22. Garrett-Jones C, Shidrawi G, 1969. Malaria vectorial capacity of a population of *Anopheles gambiae*: an exercise in epidemiological entomology. *Bull World Health Organ 40:* 531–545.
23. Scoglio CM, Bosca C, Riad MH, Sahneh FD, Britch SC, Cohnstaedt LW, Linthicum KJ, 2016. Biologically informed individual- based network model for rift valley fever in the us and evaluation of mitigation strategies. *PLoS One 11:* e0162759.
24. Riad MH, Scoglio CM, Scott McVey D, Cohnstaedt LW, 2016. An individual-level network model for a hypothetical outbreak of Japanese encephalitis in the USA. *Stochastic Environ Res Risk Assess 31:* 353–367.
25. Sekamatte M, Riad MH, Tekleghiorghis T, Linthicum KJ, Britch SC, Richt JA, Gonzalez JP, Scoglio CM, 2019. Individual-based network model for rift valley fever in kabale district, Uganda. *PloS One 14:* e0202721.
26. Liu-Helmersson J, Stenlund H, Wilder-Smith A, Rocklöv J, 2014. Vectorial capacity of *Aedes aegypti*: effects of temperature and implications for global dengue epidemic potential. *PLoS One 9:* e89783.
27. Scott TW, Amerasinghe PH, Morrison AC, Lorenz LH, Clark GG, Strickman D, Kittayapong P, Edman JD, 2000. Longitudinal studies of *Aedes aegypti* (Diptera: Culicidae) in Thailand and Puerto Rico: blood feeding frequency. *J Med Entomol 37:* 89–101.
28. Briere JF, Pracros P, Le Roux AY, Pierre JS, 1999. A novel rate model of temperature-dependent development for arthropods. *Environ Entomol 28:* 22–29.
29. Lambrechts L, Paaijmans KP, Fansiri T, Carrington LB, Kramer LD, Thomas MB, Scott TW, 2011. Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *Proc Natl Acad Sci U S A 108:* 7460–7465.
30. Focks DA, Daniels E, Haile DG, Keesling JE, 1995. A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *Am J Trop Med Hyg 53:* 489–506.
31. Yang H, Macoris MLG, Galvani K, Andrighetti M, Wanderley D, 2009. Assessing the effects of temperature on the population of *Aedes aegypti*, the vector of dengue. *Epidemiol Infect 137:* 1188–1202.
32. Lega J, Brown HE, Barrera R, 2017. *Aedes aegypti* (Diptera: Culicidae) abundance model improved with relative humidity and precipitation-driven egg hatching. *J Med Entomol 54:* 1375–1384.
33. Sahneh FD, Scoglio C, Van Mieghem P, 2013. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEEACM Trans Netw 21:* 1609–1620.
34. Sahneh FD, Vajdi A, Shakeri H, Fan F, Scoglio C, 2017. GEMFsim: a stochastic simulator for the generalized epidemic modeling framework. *J Comput Sci 22:* 36–44.
35. Boguná M, Lafuerza LF, Toral R, Serrano MÁ, 2014. Simulating non-Markovian stochastic processes. *Phys Rev E 90:* 042108.
36. Masuda N, Rocha LE, 2018. A Gillespie algorithm for non-Markovian stochastic processes. *SIAM Rev 60:* 95–115.
37. Petersen LR, Tomashek KM, Biggerstaff BJ, 2012. Estimated prevalence of dengue viremia in Puerto Rican blood donations, 1995 through 2010. *Transfusion (Paris) 52:* 1647–1651.
38. Salje H et al., 2018. Reconstruction of antibody dynamics and infection histories to evaluate dengue risk. *Nature 557:* 719–723.
39. Salje H, Lessler J, Paul KK, Azman AS, Rahman MW, Rahman M, Cummings D, Gurley ES, Cauchemez S, 2016. How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. *Proc Natl Acad Sci U S A 113:* 13420–13425.
40. Rodríguez-Barraquer I, Buathong R, Iamsirithaworn S, Nisalak A, Lessler J, Jarman RG, Gibbons RV, Cummings DA, 2013. Revisiting Rayong: shifting seroprofiles of dengue in Thailand and their implications for transmission and control. *Am J Epidemiol 179:* 353–360.
41. Rodríguez-Barraquer I, Solomon SS, Kuganantham P, Srikrishnan AK, Vasudevan CK, Iqbal SH, Balakrishnan P, Solomon S, Mehta SH, Cummings DA, 2015. The hidden burden of dengue and chikungunya in Chennai, India. *PLoS Negl Trop Dis 9:* e0003906.
42. Rahman M, Rahman K, Siddque A, Shoma S, Kamal A, Ali K, Nisaluk A, Breiman RF, 2002. First outbreak of dengue hemorrhagic fever, Bangladesh. *Emerg Infect Dis 8:* 738–740.
43. Sharmin S, Viennet E, Glass K, Harley D, 2015. The emergence of dengue in Bangladesh: epidemiology, challenges and future disease risk. *Trans R Soc Trop Med Hyg 109:* 619–627.
44. Yunus EB, Bangali AM, Mahmood MAH, Mushfiqur M, Rahman A, Talukder K, 2001. Dengue outbreak 2000 in Bangladesh: from speculation to reality and exercises. *Dengue Bull 25:* 15.
45. Bhatt S et al., 2013. The global distribution and burden of dengue. *Nature 496:* 504–507.
46. Undurraga EA, Halasa YA, Shepard DS, 2013. Use of expansion factors to estimate the burden of dengue in southeast Asia: a systematic analysis. *PLoS Negl Trop Dis 7:* e2056.
47. Dhar-Chowdhury P, Paul KK, Haque CE, Hossain S, Lindsay LR, Dibernardo A, Brooks WA, Drebot MA, 2017. Dengue seroprevalence, seroconversion and risk factors in Dhaka, Bangladesh. *PLoS Negl Trop Dis 11:* e0005475.
48. Zino L, Rizzo A, Porfiri M, 2017. An analytical framework for the study of epidemic models on activity driven networks. *J Complex Netw 5:* 924–952.
49. Chan M, Johansson MA, 2012. The incubation periods of dengue viruses. *PLoS One 7:* e50972.
50. Internet, 2020. *Climate Data for Cities Worldwide*. Available at: https://en.climate-data.org. Accessed February 12, 2020.
51. Internet, 2020. *City Population*. Available at: https://www.citypopulation.de/. Accessed February 12, 2020.

52. Sharmin S, Glass K, Viennet E, Harley D, 2015. Interaction of mean temperature and daily fluctuation influences dengue incidence in Dhaka, Bangladesh. *PLoS Negl Trop Dis 9:* e0003901.
53. Yung CF, Lee KS, Thein TL, Tan LK, Gan VC, Wong JG, Lye DC, Ng LC, Leo YS, 2015. Dengue serotype-specific differences in clinical manifestation, laboratory parameters and risk of severe disease in adults, Singapore. *Am J Trop Med Hyg 92:* 999–1005.
54. Akram A, 2019. Alarming turn of dengue fever in Dhaka city in 2019. *Bangladesh J Infect Dis 6:* 1–2.
55. Shirin T, Muraduzzaman A, Alam A, Sultana S, Siddiqua M, Khan M, Akram A, Sharif A, Hossain S, Flora M, 2019. Largest dengue outbreak of the decade with high fatality may be due to reemergence of DEN-3 serotype in Dhaka, Bangladesh, necessitating immediate public health attention. *New Microbes New Infect 29:* 100511.
56. Muraduzzaman A, Alam AN, Sultana S, Siddiqua M, Khan MH, Akram A, Haque F, Flora MS, Shirin T, 2018. Circulating dengue virus serotypes in Bangladesh from 2013 to 2016. *Virusdisease 29:* 303–307.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2021, accepted November 5, 2021, date of publication November 23, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129997

SEIR-SEI-EnKF: A New Model for Estimating and Forecasting Dengue Outbreak Dynamics

CHUNLIN YI[1], LEE W. COHNSTAEDT[2], AND CATERINA M. SCOGLIO[1], (Senior Member, IEEE)
[1]Department of Electrical and Computer Engineering, College of Engineering, Kansas State University, Manhattan, KS 66506, USA
[2]Foreign Arthropod-Borne Animal Diseases Research Unit, National Bio and Agrodefense Facility, USDA ARS, Manhattan, KS 66502, USA

Corresponding author: Lee W. Cohnstaedt (lee.cohnstaedt@usda.gov)

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 Dengue fever is an acute mosquito-borne viral infection that results in a heavy social burden in many tropical and subtropical regions. Accurate forecasts of dengue outbreak allow the local health officials to take proactive action such as positioning mosquito control equipment or preparing medical resources. We developed a new model for dengue outbreak estimation and forecast that adopts the vector-borne disease model SEIR-SEI with compartments Susceptible-Exposed-Infectious-Recovered (for host) and Susceptible-Exposed-Infectious (for vector) into the ensemble Kalman filtering (EnKF) assimilation method. The SEIR-SIR-EnKF model was first validated using synthetic dengue outbreak in twin experiments. Then, the model produced good performance when applied to estimate and forecast the dengue outbreak dynamics with real historical time-series cases in 3 different cities. Furthermore, we compared the accuracy of the real-time predictions between SEIR-SEI-EnKF model, SEIR-EnKF model, and SIR-EnKF model; we found the SEIR-SEI-EnKF model had the most accurate predictions.

INDEX TERMS Ensemble Kalman filter, estimation of disease outbreak dynamics, vector-borne disease modeling, forecast of dengue cases.

I. INTRODUCTION

Dengue fever is an acute mosquito-borne viral disease that has spread rapidly in tropical regions over the past decades [1]. During large dengue outbreaks, public health officials and healthcare facilities often cannot deploy medical personnel, or administer treatment resources and emergency vector control measures in a timely and efficient manner. Therefore, an accurate incidence forecast and advance warning system of dengue outbreaks can reduce morbidity and mortality by proactively positioning resources or targeting mosquito control to high-risk areas forecasted to have increased cases. A study in 2017 showed that dengue is increasing at the highest rate (400% between 2000 and 2013) among communicable disease, which placed a heavy social, economic and health burden globally [2]. Worldwide, there is an estimation of 360 million dengue infections per year [3]. In Singapore, large dengue outbreaks occurred almost every 5-6 years since the late 1980s despite the public health efforts [4]. Dengue is endemic in Brazil and more than 11 million cases was reported from 2000 to 2016 [5]. In Taiwan,large dengue outbreaks were recorded in 1981 (about 13000 cases over 15000 residents on Liu-Chiu Islet, Pingtung), 1988 (approximate 5000 cases in Kaohsiung), 2014 (15429 cases in Kaohsiung) and 2015 (43419 cases in Tainan) [6], [7].

Recent literature has described a number of approaches to forecast infectious disease outbreaks. One class of methods are based on the mechanistic disease models which can be described by differential equations [8]–[12]. Among the mathematical models appropriate for infectious disease, the Susceptible-Infectious-Recovered (SIR) model is the simplest one. In the SIR model, individuals are divided into 3 compartments and each compartment represents a state in the disease progression. Susceptible individuals can become infected and infectious by contact with infectious

The associate editor coordinating the review of this manuscript and approving it for publication was Seung-Hyun Kong.

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.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

ones; people who recover from infection are presented by recovered [13]. Based on the SIR model, the SEIR model introduces the new compartment E (exposed) describing the state when individuals are infected but not yet infectious. Other SIR-based models are also commonly used for disease forecast, like S-E-I-H (hospitalized)-F (funeral)-R model in [9] to predict the 2014-2015 Ebola epidemics in Liberia, SIRS model for seasonal influenza prediction in [15]. For vector-born disease such as dengue fever whose infections occur alternatively between hosts (human) and vectors (mosquito), the SEIR-SEI (host-vector) model was developed and commonly used to simulate the transition dynamics of vector borne disease [16]–[19]. However, to our knowledge, there is no previous research that uses the SEIR-SEI model to estimate the model parameters and to forecast the incidence.

Another class of methods based on machine learning techniques are used to predict the number of confirmed cases during epidemics [20], [21]. In order to have an accurate assessment about the epidemic size, health officials also need to have an estimation about the number of the infectious population who are not yet confirmed and the exposed population, which cannot be achieved by machine-learning approaches because machine learning models generally cannot describe the explicit relations between the observation (confirmed cases) and estimations (infectious and exposed population). However, the mechanistic models can be used to estimate all state variables, infer the disease transition rates, describe the outbreak characteristics, and provide understanding of the epidemic dynamics.

The classic Kalman filter (KF) was designed for data assimilation problems of linear system by R.E. Kalman in 1960 [23]. An early adaption of KF is the extended Kaman filter (EKF), which can solve the nonlinear problems by linearizing the nonlinear systems. However, the usage of EKF on large-scale or severe nonlinear systems always requires high computational cost and closure approximation from neglecting higher order derivative terms of the model. The EnKF was introduced as a parameter and state estimation method by Evensen in 1994 to resolve the closure problems of EKF [24]. In the next decades, the EnKF has been widely applied to many fields, like state and parameter estimation on reservoir models [25], hydrological models [26], and doubly fed induction generators [27]. For dengue transmission, the inference of the model parameters is a difficult task because the characteristics and transmission patterns of dengue are influenced by environmental and meteorological factors and human interactions [28]. EnKF is one of the most popular approaches to estimate the key parameters of disease models. The EnKF method in conjunction with an SEIR disease model was used to produce short-term prediction of Covid-19 cases in [14], and to develop a real-time forecast system of hand-foot-and-mouth disease outbreaks in [12]. In addition, the estimation of the local female mosquito abundance [29], which plays a very important role on dengue transmission, has not been used for dengue incidence forecast.

The performance of six filtering methods were compared to forecast the seasonal influenza activity in [22]. Those methods were the basic particle filter (PF), the maximum likelihood estimation via iterated filtering (MIF), the particle Markov chain Monte Carlo (pMCMC), the ensemble Kalman filter (EnKF), the ensemble adjustment Kalman filter (EAKF), and the rank histogram filter (RHF). The EnKF, the EAKF, the RHF and the PF more accurately forecasted the outbreak than the MIF and the pMCMC because the ensemble filters (EnKF, EAKF, and RHF) and the PF are able to adjust the model parameters prior to each iteration. They concluded that the performance of PF degrades when a larger or more complex system was applied, and that the EnKF performed better than the EAKF because the EnKF is a stochastic filter while EAKF is purely deterministic.

The goal of this study was to develop a new model to estimate and forecast dengue outbreak dynamics by integrating the SEIR-SEI disease model with the EnKF assimilation method. At first, we designed twin experiments where the SEIR-SEI-EnKF model assimilated the synthetic weekly dengue cases and produced accurate estimations and predictions of states and parameters. Second, the model estimated and forecasted the dengue outbreaks in Kaohsiung Taiwan, Singapore, and Rio de Janeiro Brazil by assimilating the historical time-series incidence. The estimated and predicted cases corresponded to the observed cases and the model was able to provide estimations of the exposed and infectious human population dynamically. Furthermore, we performed a comparison on the forecast accuracy of the SEIR-SEI-EnKF, SEIR-EnKF, and SIR-EnKF models. The results showed the SEIR-SEI-EnKF model had the most accurate forecasts. In this study, the EnKF and the SEIR-SEI disease model were integrated for the first time to estimate and forecast the epidemic dynamics for vector-borne disease like dengue fever, the model was able to provide parameter and state estimation and dengue case forecast, and the model demonstrated to have more accurate forecasts than the other two commonly used models. This dengue epidemic estimation and forecast model will potentially help healthcare agencies prepare and adapt to the changing situation during dengue outbreaks.

II. METHOD

A. DATA

In this study, weekly dengue incidence data in Kaohsiung Taiwan (population 2,773,533) from epidemic week 27/2015 to 5/2016 [31], Singapore (population 5,913,813) from epedemic week 1/2020 to 50/2020 [33] and Rio de Janeiro Brazil (population 6,023,699) from epidemic week 7/2015 to 42/2015 [34] were used. In the article, we presented the estimation and forecast results in Kaohsiung Taiwan and the results in Singapore and Rio de Janeiro were presented in appendix. Since the reported dengue cases were the number of people who were diagnosed and hospitalized, we assumed that they were not able to infect others. Therefore, we used

the cumulative weekly case count as the observed recovered population in this study.

Dengue is transmitted by the bite of adult female mosquitoes. The regional number of female mosquitoes is greatly influenced by the climate conditions like precipitation and temperature. Mosquito population modeling [30] used the regional environmental data to estimate the local female mosquito abundance. The estimated female mosquito abundance of Kaohsiung was generated this way.

B. DISEASE MODEL

In order to describe the transmission of dengue virus between host and vector population groups, the SEIR-SEI disease model is used. We consider the populations involved in disease transmission, which include humans of all ages and both sexes and adult female mosquitoes. The population of each species are divided into different compartments, which are referred as state variables. Compartments for human population are 1) susceptible (S_H), defined as the population that are unexposed to the virus but can be infected by the bites from the infectious mosquitoes; 2) exposed (E_H), defined as the population that are infected but not infectious; 3) infectious (I_H); 4) recovered (R_H), defined as the population that are recovered or died. Compartments for the vector population are 1) susceptible (S_M); 2) exposed (E_M); 3) infectious (I_M). Once the mosquitoes become infectious, they are assumed to remain so until death.

Considering that the transition from the susceptible to the infection for each species depends on the biting rate of the mosquitoes and the transmission probability. The infection rates per susceptible human and susceptible mosquito are given by

$$p_H \frac{N_M}{N_H} b \frac{I_M}{N_M} = p_H b \frac{I_M}{N_H} = \beta_H \bar{I}_M \tag{1}$$

$$p_M b \frac{I_H}{N_H} = p_M b \frac{I_H}{N_H} = \beta_M \bar{I}_H \tag{2}$$

where N_H and N_M denote the human and mosquito population size respectively. The biting rate b is the average number of bites per mosquito on human per day, which is influenced by a number of factors, like climate conditions, mosquito longevity, and so on; the transmission probability p is the probability that the susceptible can be infected per infectious bite. $\beta_H = p_H b$, $\beta_M = p_M b$ are the infection rates for each species. $\bar{I}_M = \frac{I_M}{N_H}$, $\bar{I}_H = \frac{I_H}{N_H}$ are the normalized values of the infectious population.

Transitions between compartments are described by (3).

$$\begin{aligned}
\bar{S}'_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_M - \mu_H \bar{S}_H \\
\bar{E}'_H(t) &= \beta_H \bar{S}_H \bar{I}_M - (\delta_H + \mu_H)\bar{E}_H \\
\bar{I}'_H(t) &= \delta_H \bar{E}_H - (\gamma_H + \mu_H)\bar{I}_H \\
\bar{R}'_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H \\
\bar{S}'_M(t) &= \bar{A} - \beta_M \bar{S}_M \bar{I}_H - \mu_M \bar{S}_M \\
\bar{E}'_M(t) &= \beta_M \bar{S}_M \bar{I}_H - (\delta_M + \mu_M)\bar{E}_M \\
\bar{I}'_M(t) &= \delta_M \bar{E}_M - \mu_M \bar{I}_M
\end{aligned} \tag{3}$$

where λ_H denotes the birth rate for human; μ denotes the death rate for both species; δ denotes the transition rate from exposed to infectious; γ denotes the transition rate from infectious to recovered; $\bar{A}$ denotes the normalized mosquito emergence rate. In this study, the birth rate and death rate for human are approximated as 0, $\lambda_H = 0$, $\mu_H = 0$. A report from WHO shows that the incubation time of dengue on human is 4 – 10 days and the infection duration is 2 – 7 days [35], so the rates δ_H and γ_H satisfy the following distributions $\frac{1}{\delta_H} \sim N(7, 1.5)$, $\frac{1}{\gamma_H} \sim N(4.5, 1.25)$; the average lifespan of an *Aedes* mosquito in nature is two weeks and the incubation time in mosquito is about 10 days, so the death rate and incubation rate for mosquito satisfies $\frac{1}{\mu_M} \sim N(14, \sigma)$, $\frac{1}{\gamma_H} \sim N(10, \sigma)$.

The number of new cases at time t is given by $\bar{NC}_H(t) = \bar{R}_H(t) - \bar{R}_H(t-1)$.

The other two simpler disease models – SIR model and SEIR model, are often used for dengue incidence forecast. The SIR model is expressed in the following equations (4).

$$\begin{aligned}
\bar{S}'_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_H - \mu_H \bar{S}_H \\
\bar{I}'_H(t) &= \beta_H \bar{S}_H \bar{I}_H - (\gamma_H + \mu_H)\bar{I}_H \\
\bar{R}'_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H
\end{aligned} \tag{4}$$

The equations for SIER model is shown in (5).

$$\begin{aligned}
\bar{S}'_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_H - \mu_H \bar{S}_H \\
\bar{E}'_H(t) &= \beta_H \bar{S}_H \bar{I}_H - (\delta_H + \mu_H)\bar{(E)}_H \\
\bar{I}'_H(t) &= \delta_H \bar{E}_H - (\gamma_H + \mu_H)\bar{I}_H \\
\bar{R}'_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H
\end{aligned} \tag{5}$$

The variables in (4) and (5) denote the same states or parameters as they do in SEIR-SEI model. In the transition process of both models, the intervention of mosquito is completely ignored.

C. FORECAST MODEL

In order to estimate the unknown parameters and the state variables, we used an efficient data assimilation algorithm – the Ensemble Kalman Filter (EnKF). The EnKF maintains an ensemble of the state variables and the unknown parameters. Let the state variables defined as $x_k \in R^{n\times 1}$, the unknown parameters $\Phi_k \in R^{m\times 1}$, where n is the number of states variables, and m is the number of the unknown parameters. In the ensemble presentation, they can be put together in a matrix$A_k \in R^{(n+m)\times N}$, holding N ensemble member at time k

$$A_k = \begin{bmatrix} x_k^1 & x_k^2 & \dots & x_k^N \\ \phi_k^1 & \phi_k^2 & \dots & \phi_k^N \end{bmatrix} \tag{6}$$

The prior at time $k+1$ $\hat{A}_{k+1}^-$ is given by the equation (7)

$$\hat{A}_{k+1}^- = f(\hat{A}_k) + q_k, \tag{7}$$

where $\hat{A}_k$ is the posterior at time k, f is the model system, and q_k is the model noise.

In this method, the ensemble mean $\bar{\hat{A}}_k^-$ is the best estimate of the true states. Therefore, the ensemble covariance for prior at time k P_k^- is given by equation (8).

$$P_k^- = \frac{1}{N-1}(\hat{A}_k^- - \bar{\hat{A}}_k^-)(\hat{A}_k^- - \bar{\hat{A}}_k^-)^T \tag{8}$$

The ensemble presentation of measurements can be formulated as

$$Y_k = \begin{bmatrix} y_k^1 & y_k^2 & \dots & y_k^N \end{bmatrix} = y_k + \epsilon_k, \tag{9}$$

where y_k is the true measurement and ϵ_k is the perturbation vector.

$$Y_k = H_k A_k \tag{10}$$

H_k is the measurement index matrix at time k.

The measurement covariance matrix V_k at time k can be estimated as

$$V_k = \frac{\epsilon_k \epsilon_k^T}{N-1}. \tag{11}$$

The Kalman gain is given by

$$G_k = P_k^- H_k^T (H_k P_k^- H_k^T + V_k)^{-1}. \tag{12}$$

The posterior and its covariance are then updated by

$$\hat{A}_k = \hat{A}_k^- + G_k(Y_k - H_k \hat{A}_k^-), \tag{13}$$

$$P_k = \frac{1}{N-1}(\hat{A}_k - \bar{\hat{A}}_k)(\hat{A}_k - \bar{\hat{A}}_k)^T. \tag{14}$$

In this analysis, the ensemble mean of the prior $\bar{\hat{A}}_k^-$ is our forecast at time k, and the ensemble mean of the posterior $\bar{\hat{A}}_k$ is the estimation of the states and parameters.

D. TWIN EXPERIMENTS

The twin experiments is an effective method to verify the feasibility and evaluate the accuracy of parameter estimation techniques [32]. In the twin experiments, a disease model with prescribed value of parameters is simulated to produce time-series observations of state variables. Then, we take the parameters of interest as unknown ones and use the ''observed states'' as measured data to estimate those unknown parameters by EnKF. Therefore, by comparing the estimated value and the prescribed value of the parameters, we can test the correctness and evaluate the accuracy of the estimation method.

In this study, synthetic time-series of dengue epidemic dynamics and mosquito abundance are generated by an SEIR-SEI model whose parameters and initial states condition are listed in Table 1. In this simulation, the prescribed values for β_H, β_M, and $\bar{A}$ are piece-wise constant, which means their values have a step change at week 26. This action is aim to mimic the dynamic characteristics of these parameters. We add white noise (zero mean and 0.0001 variance) to the synthetic time-series data of the state $\bar{R}_H$ (total confirmed dengue cases) and the sum of $\bar{S}_M$, $\bar{E}_M$ and $\bar{I}_M$ (mosquito abundance). Then, they are input to the EnKF as the observations. The estimated parameters and states by EnKF will be compared with the synthetic model to evaluate the performance of the EnKF.

TABLE 1. Preset parameters Θ and initial conditions $\bar{x}_0$ of SEIR-SEI model.

Θ	β_H	δ_H	γ_H	$\bar{A}$	β_M	δ_M	μ_M
Values	2.5,1.5	0.7	0.75	0.2, 0.15	2.5, 1.5	0.5	0.5
$\bar{x}_0$	$\bar{S}_{H0}$	$\bar{E}_{H0}$	$\bar{I}_{H0}$	$\bar{R}_{H0}$	$\bar{S}_{M0}$	$\bar{E}_{M0}$	$\bar{I}_{M0}$
Values	0.99995	0.00002	0.00003	0	0.1	0.0005	0

E. ESTIMATIONS AND FORECASTS OF THE SEIR-SEI-EnKF MODEL

The weekly series of dengue cases (30 weeks in total) and the mosquito abundance in Kaohsiung Taiwan is used to train and test the SEIR-SEI-EnKF model. The training period is from week 1 to week 22. The system assimilates the observations during the training period to estimate the parameters and states. The predictions are produced at the end of the training period for the next 5 weeks, and the reported dengue cases from week 23 to week 27 are used to evaluate the prediction accuracy. The simulation is replicated 200 times, each model state is initialized with a positive random value at the beginning of each simulation. The average estimations and the distribution of the predictions are presented in the results.

F. REAL-TIME PREDICTIONS OF DIFFERENT DISEASE MODELS

Three disease models – SEIR-SEI, SEIR, and SIR model are applied to the EnKF separately to produce the real-time predictions of the dengue cases in Kaohsiung Taiwan. The prediction and the assimilation occur at the same time i.e. the systems produce predictions for the next 5 weeks at each week after the observation is updated. Thus, 5 steps of predictions are produced in this simulation (predictions in step 1 are the predictions for the next week, predictions in step 2 are the ones in 2 weeks, etc.). Simulations for each model is repeated 200 times. The root mean square (RMS) errors between the predictions and the corresponding observations are calculated.

III. RESULTS

A. TWIN EXPERIMENTS

Fig.1(a) shows the estimated value and prescribed value of unknown parameters. In the figure, the estimated infection rate from mosquito to human β_H converges to the prescribed value after week 13 and follows the change at week 26 (1 week delay) and finally be stable at 1.8 (20% error); the estimation of β_M has a large overshot and oscillations but it converges to the first-stage prescribed value after week 20, the EnKF detects its step decrease at week 26 and the curve converges after week 30; the estimated value of mosquito emergence rate $\bar{A}$ matches exactly with the prescribed one after week 7 and EnKF is able to detect its step change from the observation immediately.

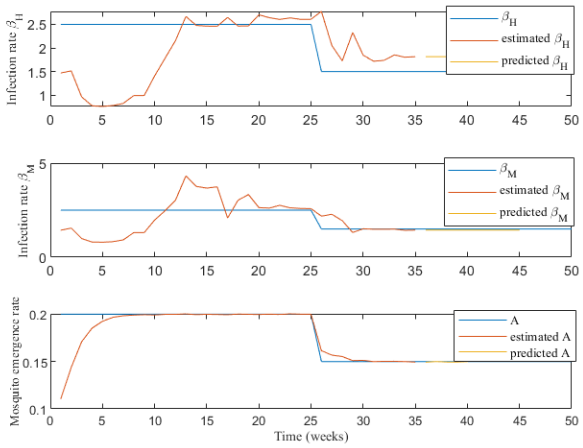
Fig.1(b)(c) show the estimated and predicted epidemic dynamics on both populations. The blue curve in each graph shows the "real" epidemic dynamics (week 1 to week 45); the red curve shows the estimated ones (week 1 to week 35); the orange curve shows the predicted one (week 36 to week 45). For human population, the value of $\bar{R}_H$ is the observation and the value of other states are unknown. As shown in Fig.1(b), the estimated and predicted values of all states are very close to their real value except the peak values of states $\bar{E}_H$ and $\bar{I}_H$. In addition, the estimated epidemic peak week (the week when $\bar{I}_H$ reaches the maximum) matches the real one. The errors for peak value of states $\bar{E}_H$ and $\bar{I}_H$ are 20% and 13.5%. For mosquito population, the total number of female mosquitos is observed. As seen in Fig.1(c), the estimations and predictions on states $\bar{S}_M$ and $\bar{E}_M$ are almost accurate. For state $\bar{I}_M$, the error is relatively large, but the estimation captures the trend of real data.

B. ESTIMATIONS AND FORECASTS OF SEIR-SEI-EnKF MODEL

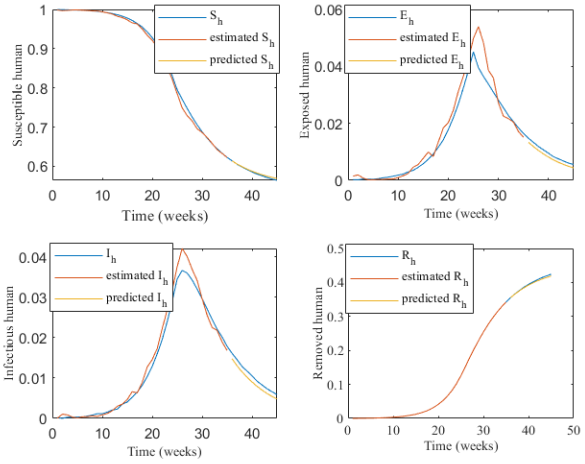
The estimated weekly infection rates of dengue for both species and the mosquito emergence rate are shown in Fig.2(a). Dengue cases was detected from the first week but the infection rates are very low in the first 10 weeks. During high-incidence period (week 10 to week 22), the infection rates for both species keep increasing to their peak value at week 21 (1.52 for β_H and 5.65 for β_M); the mosquito emergence rate A has two "steady states" – around 3.3×10^{-4} before peak weeks and around 1.8×10^{-4} after. The value of infection rate β_M is much greater than β_H, which means mosquitoes are more likely to be infected by infectious human. The estimations and the forecasts of the model states are presented in Fig.2(b)(c). On the graph, the dots are estimated values of each states; the boxes are the distribution of the predicted value over 200 repeated simulations (the red line inside the box is the median, the upper and lower edges are the quantiles); the solid line is the observed recovered human population. In Fig.2(b), the estimated recovered human population is perfectly matched the observed data. The statistics of the predictions are listed in Table 2. Taking the median as our best forecasts, the root mean square (RMS) error is 455 that is over 40 times smaller than the real values. The estimated peak week for exposed human population and infectious human population are the week 21. The estimated normalized magnitude of the infectious population is 9.8×10^{-4} that means the estimated maximum infectious population is 2715. The epidemic dynamics of mosquito population is shown in Fig.2(c). The estimated number of infectious mosquitoes is very high from week 10 to week 22, which corresponds to the high-incidence period for human population.

C. REAL-TIME PREDICTIONS OF DIFFERENT DISEASE MODELS

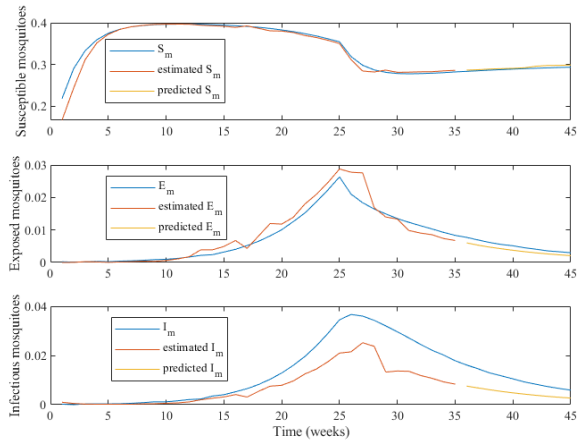
We further test the prediction performance of the SEIR-SEI-EnKF model by comparing its real-time prediction results with the SEIR-EnKF model and the SIR-EnKF model.



(a) Parameter estimation



(b) Epidemic dynamics on human population



(c) Epidemic dynamics on mosquito population

FIGURE 1. Estimation and prediction results for twin experiments.

Each model assimilate observations at each week and produce real-time predictions in the next 5 weeks.

TABLE 2. Statistics of the predictions.

Predictions	Week 23	Week 24	Week 25	Week 26	Week 27
Upper quantile	17550	18426	18926	19231	19407
Median	17469	18217	18641	18876	18978
Lower quantile	17374	18040	18373	18514	18067
Real value	17542	18439	19066	19428	19682

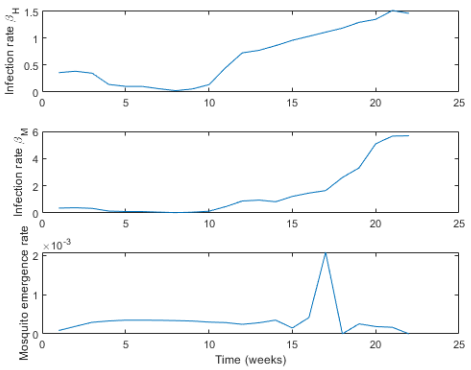
TABLE 3. Peak value of predicted new cases and their absolute relative errors (ARE).

Peak/ARE	1 step	2 step	3 step	Average ARE
SEIR-SEI	2382/7.3%	2054/20%	1725/32.9%	20.1%
SEIR	3070/19.5%	3624/41.1%	4299/67.3%	42.7%
SIR	3016/17.4%	3550/38.2%	4200/63.5%	39.7%

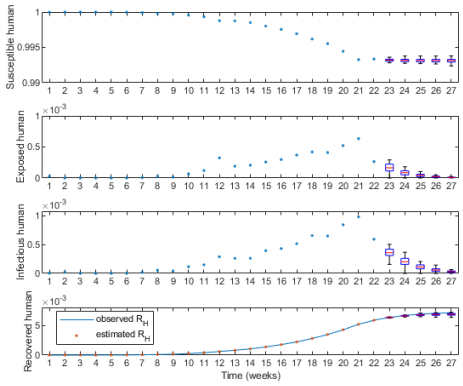
Fig.3 shows the prediction results of the total confirmed cases, which is the predicted recovered human population. The 1 step prediction refers to the predictions for the next week of the observation; the 2 step prediction are the predicted cases in 2 weeks of the observation week, etc. In the figure, dots present observations and the curves presents the predictions of different models. Overall, the first 3 predictions of all 3 models correspond well with the observations. In the step 4 prediction and the step 5 prediction, the SEIR-SEI-EnKF model underestimates the cases while the SEIR-EnKF model and SIR-EnKF model overestimate that.

The real-time predictions of the weekly new confirmed cases are presented in Fig.4. The new confirmed cases is given by $NC_H^i = R_H^i - R_H^{i-1}$, which is not a state of the disease model. Meanwhile, the reported weekly cases is not the model observation neither. Apparently, the relative errors of the weekly case predictions are greater than that of the confirmed case, because the EnKF adjusts the model parameters and calculates the Kalman gain by the error between the prior and the corresponding observation which is the total confirmed cases. The step 4 and the step 5 predictions of each model are not able to reflect the weekly new confirmed cases. The peak week of the reported weekly confirmed cases is week 21; the predicted peak weeks of SEIR-SEI-EnKF model are week 22 for step 1, week 22 for step 2, and week 23 for step 3; those of SEIR-EnKF model and SIR-EnKF model are week 22 for step 1, week 23 for step 2, and week 24 for step 3. The peak value of the reported new confirmed cases is 2569 and Table 3 lists the predicted peak values of new confirmed cases for each model and their absolute relative errors (ARE) with the reported one. The results show that the models are able to detect the peak of the new confirmed cases by observing the total confirmed cases and that the SEIR-SEI-EnKF model has more accurate performance on predicting the peak time and peak value of the weekly confirmed cases.

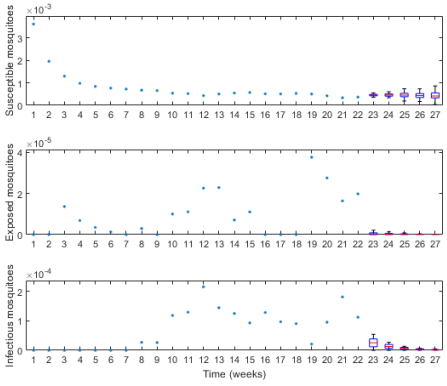
To evaluate the prediction performance more objectively, we calculate the RMS error of each prediction. Fig.5 shows the RMS errors over each prediction step for each model and there are three evident patterns: (1) the 1 step prediction errors for all 3 models are very close, which means that all



(a) Parameter estimation



(b) Epidemic dynamics on human population



(c) Epidemic dynamics on mosquito population

FIGURE 2. Estimation and prediction results for SEIR-SEI-EnKF model.

3I models are able to have very good performance when we only forecast the cases in the next week; (2) for each model, the errors over the earlier prediction steps are smaller; (3) the prediction errors for SEIR-SEI-EnKF model are smaller than those for SEIR-EnKF model and SIR-EnKF model. Moreover, the errors for SIR-EnKF model increase exponentially with the prediction weeks. Therefore, when we try to forecast the dengue cases in the long term, the SIR model is not robust

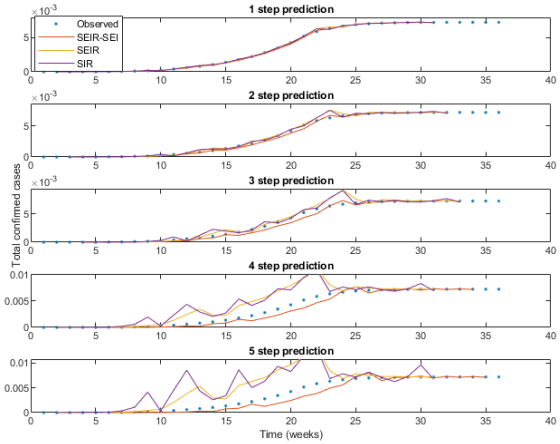


FIGURE 3. Real-time predictions of total confirmed cases.

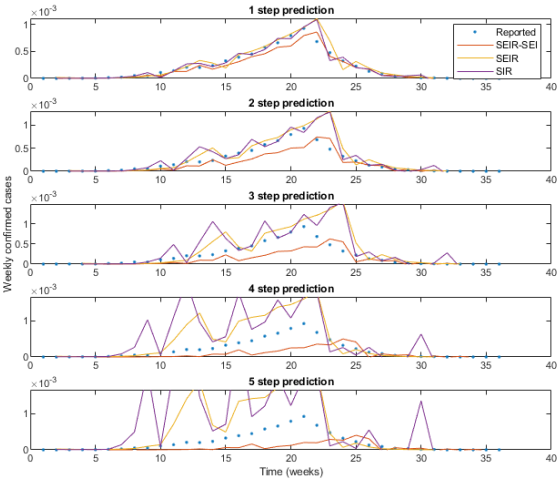


FIGURE 4. Real-time predictions of new confirmed cases.

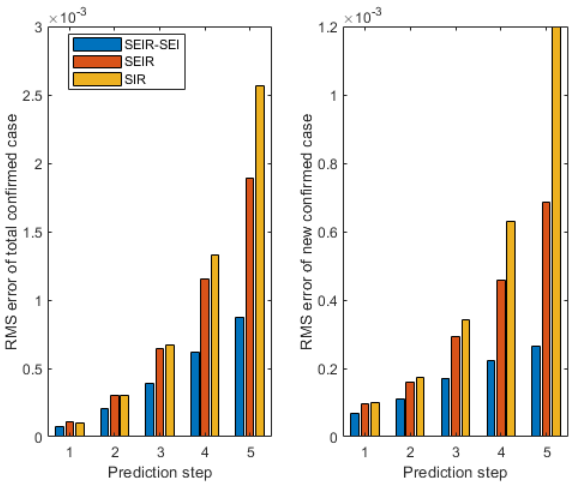
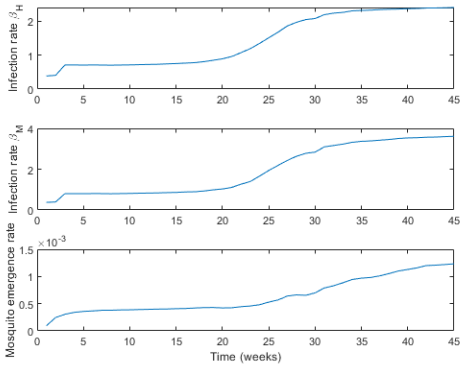
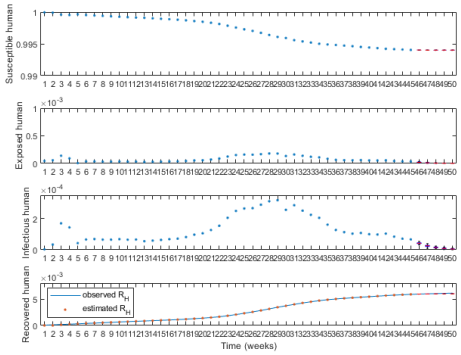


FIGURE 5. RMS errors of real-time predictions.

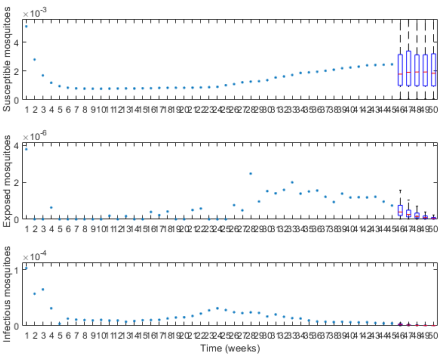
and reliable. The prediction errors from SEIR-EnKF model are smaller than SIR model but greater than SEIR-SEI model, and they increase faster than SEIR-SEI-EnKF model.



(a) Parameter estimation



(b) Epidemic dynamics on human population



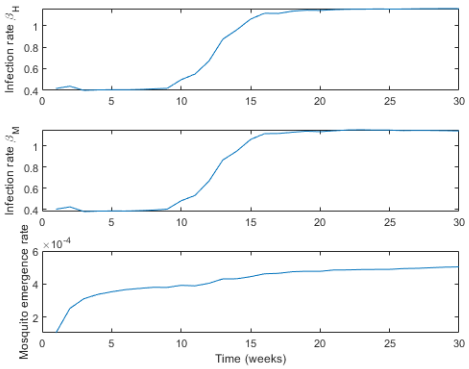
(c) Epidemic dynamics on mosquito population

FIGURE 6. Estimation and prediction results for 2020 dengue outbreak in Singapore.

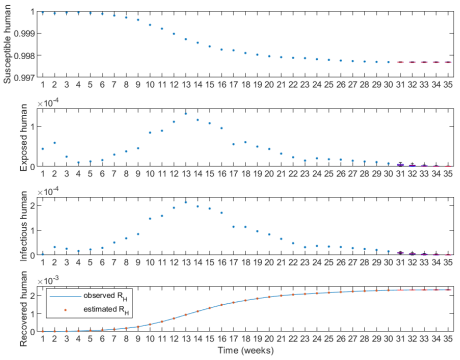
IV. DISCUSSION

We developed a new SEIR-SEI-EnKF model for estimating and forecasting dengue outbreak dynamics. The model estimated the model parameters and states, and dynamically produced accurate short-term forecasts by assimilating the observations.

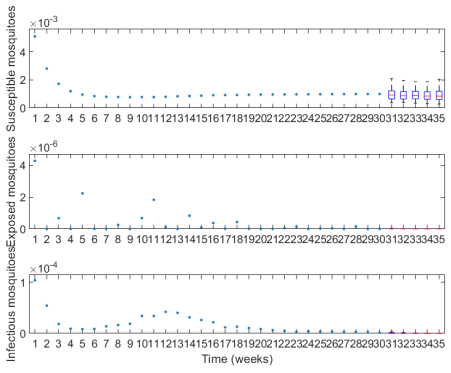
Dengue fever is a mosquito-borne viral disease that is dependent on the mosquito and human interactions. Consequently, high mosquito populations are correlated with elevated viral transmission and infection rates. In the disease



(a) Parameter estimation

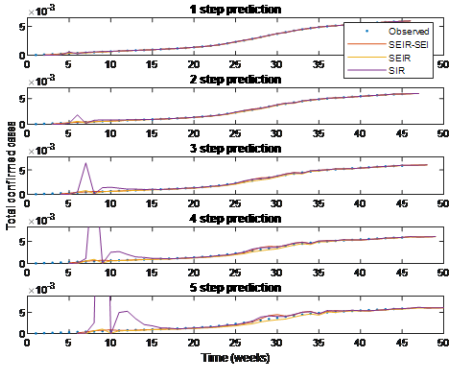


(b) Epidemic dynamics on human population

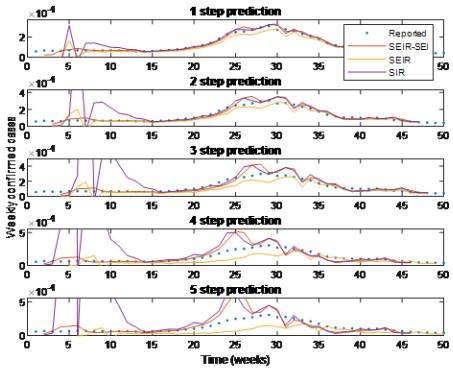


(c) Epidemic dynamics on mosquito population

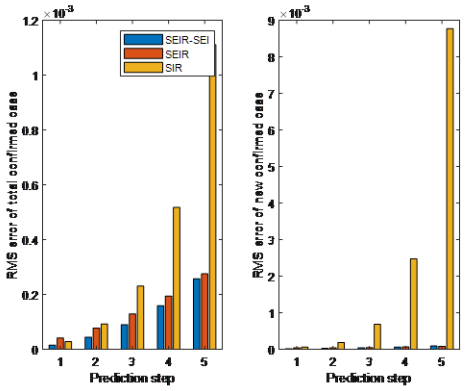
FIGURE 7. **Estimation and prediction results for 2015 dengue outbreak in RDJ.**



(a) Real-time predictions for total confirmed cases



(b) Real-time predictions for new confirmed cases

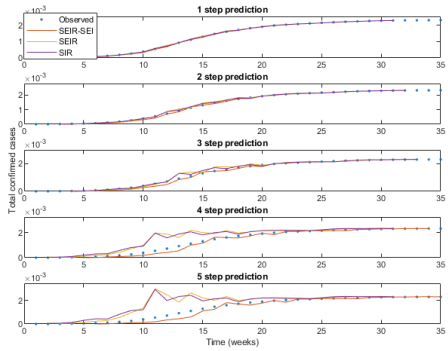


(c) RMS errors of different models

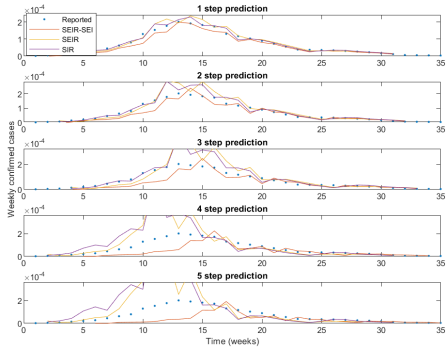
FIGURE 8. **Real-time predictions for 2020 dengue outbreak in Singapore.**

transmission, the infection rates and the mosquito abundance are essential factors which have great influence on the overall epidemic size. Weather conditions, such as warmer temperature and increased precipitation, are correlated to increased viral transmission, whereas human interventions (mosquito management) reduce contact with mosquitoes which reduces transmission. In our disease model, the three estimated parameters (two infection rates β_H, β_M and the mosquito emergence rate A) were taken as small positive random constants initially and adjusted during assimilation by EnKF. The dependence of these parameters on weather added error to the estimations if local weather data was not available.

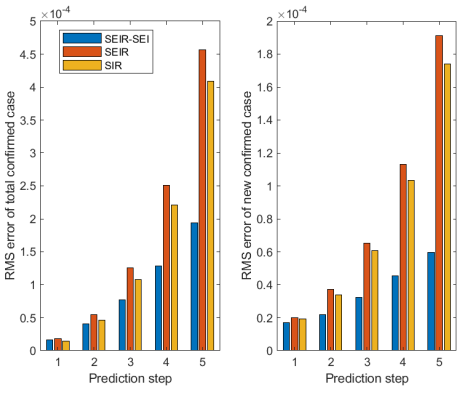
The SEIR-SEI model generated more accurate forecasts because it accounted for the mosquito population when forecasting the dengue transmission. Another reason for the increased accuracy was that the SEIR-SEI model had a higher degree of freedom than the other two models; thus, the SEIR-SEI model had more flexibility to fit the data.



(a) Real-time predictions for total confirmed cases



(b) Real-time predictions for new confirmed cases



(c) RMS errors of different models

FIGURE 9. Real-time predictions for 2015 dengue outbreak in RDJ.

Higher degree systems become over-fitted when more observations are fed into the systems. Our forecasts in three cities demonstrate SEIR-SEI model is the better-fitted one than SIR and SEIR models.

Based on the simulations on the synthetic data and the simulations on the historical time-serial data, the SEIR-SEI-EnKF model is reliable at estimating and predicting dengue outbreak dynamics. During an on-going dengue outbreak, this model is able to produce incidence predictions and estimate the current exposed and infectious populations, which is required for the public health agencies to optimize the treatment resources and take control measures in advance.

There are several limitations in our study. First, the collected dengue case data did not include the infected population who did not seek for medical assistance, which rendered the reported dengue cases was the lower bond of the actual case. Further, the estimated parameters (two infection rates β_H, β_M and the mosquito emergence rate A) might be better estimated using available weather data and this will be one of our limitations and future work. This study did not compare all possible adaptions of disease models; there might be an adaption of SEIR-SEI model which has more accurate forecast than SEIR-SEI model.

APPENDIX

A. ESTIMATIONS AND FORECASTS OF SEIR-SEI-EnKF MODEL FOR SINGAPORE AND RDJ

Fig.6 and Fig.7 show the parameter estimations, estimated and predicted dengue outbreak dynamics in Singapore and Rio de Janeiro Brazil based on the SEIR-SEI-EnKF model. The dengue case data from epidemic week 1/2020 to 50/2020 in Singapore and the case data from epidemic week 7/2015 to 42/2015 were used as the observations. The simulation results are similar with the ones in Kaohsiung Taiwan. The curves for the infection rates are similar and the value of $beta_M$ is larger; the estimated and predicted recovered population corresponds the actual data. The epidemic peak time is estimated at week 29 in Singapore and week 13 in RDJ.

B. REAL-TIME PREDICTION OF DIFFERENT DISEASE MODELS FOR SINGAPORE AND RDJ

Fig.8 and Fig.9 show the real-time predictions of the dengue cases in Singapore and RDJ from three models. The results are also similar with the ones in Kaohsiung. The 1 step predictions for each model are the most accurate among all 5 steps ones, and the predictions from SEIR-SEI model are more accurate than the other two models.

ACKNOWLEDGMENT

The authors would like to thank Dr. Aram Vajdi for providing mosquito population data. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

REFERENCES

[1] M. G. Guzman, "A new moment for facing dengue?" *Pathogens Global Health*, vol. 109, no. 1, pp. 2–3, Feb. 2015.

[2] C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, and T. Fleming, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study," *JAMA Oncol.*, vol. 3, no. 4, pp. 524–548, 2017.

[3] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, and M. F. Myers, "The global distribution and burden of dengue," *Nature*, vol. 496, no. 7446, pp. 504–507, 2013.

[4] J. Rajarethinam, L. W. Ang, J. Ong, and J. Ycasas, "Dengue in Singapore from 2004 to 2016: Cyclical epidemic patterns dominated by serotypes 1 and 2," *Amer. J. Tropical Med. Hygiene*, vol. 99, no. 1, p. 204, 2018.
[5] E. J. A. Luna, G. M. Figueiredo, J. E. Levi, S. R. S. Campos, A. C. Felix, N. S. E. Souza, W. M. Figueiredo, A. A. Costa, M. R. A. Cardoso, and C. S. Pannuti, "A cohort study to assess the incidence of dengue, Brazil, 2014–2018," *Acta Tropica*, vol. 204, Apr. 2020, Art. no. 105313.
[6] W.-J. Chen, "Dengue outbreaks and the geographic distribution of dengue vectors in taiwan: A 20-year epidemiological analysis," *Biomed. J.*, vol. 41, no. 5, pp. 283–289, Oct. 2018.
[7] J.-J. Tsai, C.-K. Liu, W.-Y. Tsai, L.-T. Liu, J. Tyson, C.-Y. Tsai, P.-C. Lin, and W.-K. Wang, "Seroprevalence of dengue virus in two districts of Kaohsiung city after the largest dengue outbreak in Taiwan since world war II," *PLOS Neglected Tropical Diseases*, vol. 12, no. 10, Oct. 2018, Art. no. e0006879.
[8] T. K. Yamana, S. Kandula, and J. Shaman, "Superensemble forecasts of dengue outbreaks," *J. Roy. SoC. Interface*, vol. 13, no. 123, Oct. 2016, Art. no. 20160410.
[9] S. Venkatramanan, B. Lewis, J. Chen, D. Higdon, A. Vullikanti, and M. Marathe, "Using data-driven agent-based models for forecasting emerging infectious diseases," *Epidemics*, vol. 22, pp. 43–49, Mar. 2018.
[10] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani, "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt," *Epidemics*, vol. 22, pp. 13–21, Mar. 2018.
[11] P. Narula, V. Piratla, A. Bansal, S. Azad, and P. Lio, "Parameter estimation of tuberculosis transmission model using ensemble Kalman filter across Indian states and union territories," *Infection, Disease Health*, vol. 21, no. 4, pp. 184–191, Dec. 2016.
[12] Z. Zhan, W. Dong, Y. Lu, P. Yang, Q. Wang, and P. Jia, "Real-time forecasting of hand-foot-and-mouth disease outbreaks using the integrating compartment model and assimilation filtering," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.
[13] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. London, U.K.: Oxford Univ. Press, 1992.
[14] Q. Yang, C. Yi, A. Vajdi, L. W. Cohnstaedt, H. Wu, X. Guo, and C. M. Scoglio, "Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China," *Infectious Disease Model.*, vol. 5, pp. 563–574, 2020.
[15] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 50, pp. 20425–20430, Dec. 2012.
[16] L. Esteva and C. Vargas, "Analysis of a dengue disease transmission model," *Math. Biosci.*, vol. 150, no. 2, pp. 131–151, Jun. 1998.
[17] M. Andraud, N. Hens, C. Marais, and P. Beutels, "Dynamic epidemiological models for dengue transmission: A systematic review of structural approaches," *PLoS ONE*, vol. 7, no. 11, Nov. 2012, Art. no. e49085.
[18] M. Otero, D. H. Barmak, C. O. Dorso, H. G. Solari, and M. A. Natiello, "Modeling dengue outbreaks," *Math. Biosciences*, vol. 232, no. 2, pp. 87–95, Aug. 2011.
[19] R. A. Erickson, S. M. Presley, L. J. S. Allen, K. R. Long, and S. B. Cox, "A dengue model with a dynamic Aedes albopictus vector population," *Ecol. Model.*, vol. 221, no. 24, pp. 2899–2908, Dec. 2010.
[20] F. Liang, P. Guan, W. Wu, and D. Huang, "Forecasting influenza epidemics by integrating internet search queries and traditional surveillance data with the support vector machine regression model in Liaoning, from 2011 to 2015," *PeerJ*, vol. 6, p. e5134, Jun. 2018.
[21] H. Hu, H. Wang, F. Wang, D. Langley, A. Avram, and M. Liu, "Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 4895.
[22] W. Yang, A. Karspeck, and J. Shaman, "Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics," *PLoS Comput. Biol.*, vol. 10, no. 4, Apr. 2014, Art. no. e1003583.
[23] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.
[24] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," *J. Geophys. Res.*, vol. 99, no. 5, pp. 10143–10162, 1994.
[25] J. P. Jensen, "Ensemble Kalman filtering for state and parameter estimation on a reservoir model," M.S. thesis, Dept. Eng. Cybern., Norwegian Univ. Sci. Technol., Trondheim, Norway, 2007.
[26] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser, "Dual state–parameter estimation of hydrological models using ensemble Kalman filter," *Adv. Water Resour.*, vol. 28, no. 2, pp. 135–147, Feb. 2005.
[27] R. Fan, Z. Huang, S. Wang, R. Diao, and D. Meng, "Dynamic state estimation and parameter calibration of a DFIG using the ensemble Kalman filter," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Dec. 2015, pp. 1–5.
[28] C. W. Morin, A. J. Monaghan, M. H. Hayden, R. Barrera, and K. Ernst, "Meteorologically driven simulations of dengue epidemics in san juan, PR," *PLOS Neglected Tropical Diseases*, vol. 9, no. 8, Aug. 2015, Art. no. e0004002.
[29] H. M. Yang, M. L. G. Macoris, K. C. Galvani, M. T. M. Andrighetti, and D. M. V. Wanderley, "Assessing the effects of temperature on dengue transmission," *Epidemiol. Infection*, vol. 137, no. 8, pp. 1179–1187, Aug. 2009.
[30] A. Vajdi, "Personal communication," Netw. Sci. Eng. Group, Dept. Elect. Comput. Eng., College Eng., Kansas State Univ., Manhattan, KS, USA, Private Commun., Feb. 2021.
[31] *Taiwan National Infectious Disease Statistics System*. Accessed: Feb. 6, 2016. [Online]. Available: https://nidss.cdc.gov.tw/en/nndss/disease?id=061
[32] K. Sasaki, A. Yamanaka, S.-I. Ito, and H. Nagao, "Data assimilation for phase-field models based on the ensemble Kalman filter," *Comput. Mater. Sci.*, vol. 141, pp. 141–152, Jan. 2018.
[33] *Singapore National Environment Agency*. Accessed: Dec. 12, 2020. [Online]. Available: https://www.nea.gov.sg/dengue-zika/dengue/dengue-cases
[34] 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.
[35] *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control. Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*, World Health Organization, Geneva, Switzerland, 2009, pp. 1–147.



CHUNLIN YI received the B.S. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Kansas State University under the supervision of Dr. Caterina M. Scoglio. His research interests include network science, modeling of epidemics, and Kalman filtering.

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, 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 "Sapienza" Rome University, Italy, in 1987. Before joining Kansas State University, she was with the Fondazione Ugo Bordoni, from 1987 to 2000, and the Georgia Institute of Technology, from 2000 to 2005. She is currently the Paslay Chair Professor of Electrical and Computer Engineering with Kansas State University. Her main research interests include network science and engineering.

Infectious Disease Modelling 5 (2020) 563–574

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China

Qihui Yang [a, *, 1], Chunlin Yi [a, 1], Aram Vajdi [a], Lee W. Cohnstaedt [b], Hongyu Wu [a], Xiaolong Guo [a], Caterina M. Scoglio [a]

[a] Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS, USA
[b] Arthropod-Borne Animal Diseases Research Unit, Center for Grain and Animal Health Research, USDA ARS, Manhattan, KS, USA

ARTICLE INFO

Article history:
Received 27 March 2020
Received in revised form 17 June 2020
Accepted 5 August 2020
Handling editor: Dr. J Wu

Keywords:
COVID-19
Kalman filtering
Network-based model
Non-Markovian process

ABSTRACT

As an emerging infectious disease, the 2019 coronavirus disease (COVID-19) has developed into a global pandemic. During the initial spreading of the virus in China, we demonstrated the ensemble Kalman filter performed well as a short-term predictor of the daily cases reported in Wuhan City. Second, we used an individual-level network-based model to reconstruct the epidemic dynamics in Hubei Province and examine the effectiveness of non-pharmaceutical interventions on the epidemic spreading with various scenarios. Our simulation results show that without continued control measures, the epidemic in Hubei Province could have become persistent. Only by continuing to decrease the infection rate through 1) protective measures and 2) social distancing can the actual epidemic trajectory that happened in Hubei Province be reconstructed in simulation. Finally, we simulate the COVID-19 transmission with non-Markovian processes and show how these models produce different epidemic trajectories, compared to those obtained with Markov processes. Since recent studies show that COVID-19 epidemiological parameters do not follow exponential distributions leading to Markov processes, future works need to focus on non-Markovian models to better capture the COVID-19 spreading trajectories. In addition, shortening the infectious period via early case identification and isolation can slow the epidemic spreading significantly.

© 2020 The Authors. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The 2019 novel coronavirus disease (COVID-19) emerged in Wuhan City, Hubei Province of China, in the late fall of 2019 (Chen et al., 2020). The highly contagious virus outbreak during the Chinese Spring festival period made containment an unprecedented challenge. The risks of human-to-human transmission, long persistence on surfaces, and asymptomatic infections increased spread and made case detection difficult (Chan et al., 2020; Rothe et al., 2020). Increasing case reports saturated hospitals in Hubei Province, overwhelming medical resources. Despite the high case reports, the true infection rate

* Corresponding author.
E-mail address: qihui@ksu.edu (Q. Yang).
Peer review under responsibility of KeAi Communications Co., Ltd.
[1] These authors equally contributed.

https://doi.org/10.1016/j.idm.2020.08.001
2468-0427/© 2020 The Authors. Production and hosting by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

was likely underreported, as infected individuals with only mild symptoms might have recovered without being detected (Nishiura et al., 2020; Sanche et al., 2020).

To stop the viral spread, a series of strict control measures have been implemented in mainland China, including the enforcement of the tourism ban, suspension of the school, and extension of the Spring Festival holiday. From January 23rd to April 8th, 2020, the Chinese government implemented a metropolitan-wide quarantine of Wuhan and nearby cities, suspending all public transportation and advising residents to stay at home (Du et al., 2020; Wu, Leung, & Leung, 2020). Since February 8th, a door-to-door check was started in Wuhan to take all confirmed and suspected people into medical care, which ended at around February 19th (Zhao et al., 2020). Since February 10th, Wuhan City started to close the community.

Understanding the impact of non-pharmaceutical interventions on the epidemic dynamics is crucial for decision-makers to take proper actions to contain the COVID-19. Since the beginning of the outbreak, scientists have been actively retrieving the epidemiological characteristics such as the time from infection to illness onset (incubation period) and the time from illness onset to hospital admission or isolation (infectious period) (Chen et al., 2020; Guan et al., 2020; Huang et al., 2020). Researchers have used the reported cases during the early stage of the epidemic and predicted the epidemic trajectory, the basic reproductive number R_0 from which ranges from 2.2 to 6.6 (Read, Bridgen, Cummings, Ho, & Jewell, 2020; Sanche et al., 2020; Zhao et al., 2020). Several studies explored the effects of the quarantine of Wuhan (Li, Zhao, & Sun, 2020; Shen, Peng, Guo, Xiao, & Zhang, 2020), travel restrictions (Chinazzi et al., 2020), and other non-pharmaceutical intervention strategies (Lai et al., 2020; Zhu et al., 2020) on the future transmission dynamics at different geographical scales.

China's aggressive and prompt measures have led to a dramatic decrease in the number of new cases. Those measures saved many lives, although at a high economic cost. As of March 18th, 2020, the cumulative number of confirmed cases in China was 81,186. Compared to the previous day, the number of new cases was only 70, most of which were imported from outside the country (Wu, Ge, Yu, & Hu, 2020). In the meantime, COVID-19 spread further around the world, and parts of Italy, South Korea, the United States, and Iran struggled to contain the epidemic.

In this work, we aim to assess the capability of the ensemble Kalman filter as a good short-term predictor, and test the effectiveness of non-pharmaceutical interventions on the epidemic spreading. First, we provide real-time assessments and forecasts of the case reported in Wuhan City based on the ensemble Kalman filter. Second, we build an individual-level based network model and perform stochastic simulations to reconstruct the epidemics in Hubei Province at its early stage and examine the epidemic dynamics under different scenarios. We consider four scenarios: keeping the early-stage trend without any mitigation, reducing the average node degree through social distancing, reducing the infection rate by adopting protective measures, and decreasing the infection rate through both social distancing and protective measures. Third, we simulate the epidemic spreading in Wuhan City incorporating non-Markovian processes, i.e., non-exponential parameter distributions, and compare the predicted epidemic trajectories obtained from Markov processes.

Mathematical modeling can forecast the size and timing of epidemics, so resource allocations and decision-making can be optimized to maximize intervention strategies. Since most current network-based models are based on Markov processes, this work is probably one of the first to simulate the COVID-19 spreading based on non-Markovian processes. The findings of this work can be a reference for policy recommendations regarding COVID-19 in other countries.

2. Model

In this section, we first introduce the ensemble Kalman filtering method used in the short-term prediction. Then we illustrate epidemic models based on Markov processes and non-Markovian processes built in the long-term analysis.

To describe the epidemiology of COVID-19 pneumonia, we use the Susceptible-Exposed-Infectious-Removed (SEIR) compartmental model, which has been frequently used in the study of the COVID-19 (Li, Zhao, & Sun, 2020; Read, Bridgen, Cummings, Ho, & Jewell, 2020; Wu, Leung, & Leung, 2020; Zhao et al., 2020). Each node is either in one of the compartments susceptible, exposed, infectious, or removed. Susceptible nodes refer to people free of the virus and can become infected due to contact with infectious nodes. Exposed nodes have been infected by the disease, but have not yet shown symptoms or become infectious. Removed nodes represent those individuals who are recovered, isolated, hospitalized, or dead due to the disease.

Right after the COVID-19 outbreak in China, many studies have estimated the epidemiological parameters, which are essential to calibrate the epidemic models. However, we noticed that there exists a variety regarding the epidemiological parameters. Based on 1099 patients in 31 provincial municipalities through January 29th, 2020, Guan et al. (2020) estimated that the median incubation period was 4.0 days (interquartile range 2–7 days). Based on a smaller sample size of 425 patients, Li, Guan, et al. (2020) estimated the mean incubation period to be 5.2 days (95% confidence interval, 4.1–7.0 days). Du et al. (2020) considered the delay from symptom onset to case detection as 4–5 days. Considering the possibility of asymptomatic infections, we adopt a relatively short mean incubation (latent) period as 3 days and mean infectious period as 4 days in the short-term prediction and long-term analysis based on Markovian processes (Lin et al., 2020).

2.1. Short-term prediction based on Kalman filtering

Short-term prediction is provided for the COVID-19 spreading in Wuhan city using the Kalman filtering, which consists of model state estimation at each observation time based only on the observation up to the present. Considering that there is an unknown time-varying parameter β in the SEIR model which cannot be dealt with by widely used classic Kalman filter

(Kalman, 1960) and extended Kalman filter (Jazwinski, 2007), a dual state-parameter estimation filtering method—ensemble Kalman filter (EnKF) is used for our study (Jensen, 2007; Moradkhani, Sorooshian, Gupta, & Houser, 2005). EnKF is based on Monte Carlo, where the approximation of forecast (a priori) state error covariance matrix is made by propagating an ensemble of model states using the updated states (ensemble members) from the previous time step.

We use a generalized network, in which every individual is connected with a possibility. At any discrete time, every individual has to be in one of the four compartments in the SEIR compartmental model. The equations for the state transitions are shown as follows.

$$S_{k+1} = S_k - \beta S_k I_k \Delta T \tag{1}$$

$$E_{k+1} = E_k + \beta S_k I_k \Delta T - \lambda E_k \Delta T \tag{2}$$

$$I_{k+1} = I_k + \lambda E_k \Delta T - \delta I_k \Delta T \tag{3}$$

$$R_{k+1} = R_k + \delta I_k \Delta T \tag{4}$$

where β is the infection rate; λ is the transition rate from the exposed compartment to the infectious compartment; δ denotes the removal rate. $\Delta T = 1$ day and $k = 0,\ 1,\ 2, 3$ is the k^{th} time step. We assume that the infection process of each infectious individual is Poisson and statistically independent, and that the removing process for every infected individual is also Poisson. The infection rate β quantifies the probability that each susceptible individual can be infected by each infectious individual in a time unit. The rate λ quantifies the average portion of the exposed population that will become infectious in a time unit. The removal rate δ quantifies the average portion of the infectious population that will clear the infection in a time unit. Considering that the population flow, prevention measures, and self-quarantine could affect the infection rate, we take the infection rate β as a time-varying parameter. The other two rates λ and δ are set as constants 1/3 day^{-1} and 1/4 day^{-1} (Lin et al., 2020).

Let the state variables (SEIR) defined as $x_k \in R^4$, the time-varying parameter $\beta_k \in R$. Consider the following problem

$$x_{k+1} = f(x_k, \beta_k) + q_k \tag{5}$$

$$\beta_{k+1} = h(\beta_k) \tag{6}$$

$$y_k = g(x_k, \beta_k) + v_k \tag{7}$$

where q_k and v_k are model and observation noise which are assumed to be normally distributed and with zero mean and the covariance known.

As the true states are generally unknown, we use the ensemble mean as the best estimate of the true state. The error covariance matrices for the predicted and filtered ensemble are given by

$$P_k^- = E\left[\left(\widehat{x}_k^- - E\left[\widehat{x}_k^-\right]\right)\left(\widehat{x}_k^- - E\left[\widehat{x}_k^-\right]\right)^T\right] \tag{8}$$

$$P_k = E\left[\left(\widehat{x}_k - E[\widehat{x}_k]\right)\left(\widehat{x}_k - E[\widehat{x}_k]\right)^T\right] \tag{9}$$

where P_k^- is the error covariance matrix associated with the forecasted estimate (prior), P_k is the one associated with the updated estimate (posterior), $\widehat{x}_k^-$ is the ensemble vector of the forecasted estimate from model equations.

In the ensemble presentation, the model states x_k and the time-varying parameter β_k can be put together in a matrix $A_k \in R^{5\times N}$, holding N ensemble member at time k.

$$A_k = \begin{bmatrix} x_k^1 & x_k^2 & \dots & x_k^N \\ \beta_k^1 & \beta_k^2 & \dots & \beta_k^N \end{bmatrix} \tag{10}$$

The ensemble covariance then can be presented as

$$P_k = \frac{1}{N-1}\left(\widehat{A}_k - \overline{\widehat{A}}_k\right)\left(\widehat{A}_k - \overline{\widehat{A}}_k\right)^T \tag{11}$$

$$\overline{A}_k = \frac{1}{N} A_k \cdot 1_{N\times 1} \tag{12}$$

The measurement can also be formulated in an ensemble presentation

$$Y_k = \left[y_k^1\ y_k^2\ \ldots\ y_k^N \right] \tag{13}$$

$$\varepsilon_k = \left[v_k^1\ v_k^2\ \ldots\ v_k^N \right] \tag{14}$$

where $y_k^1 = y_k + v_k^1$, y_k is the true measurement and v_k^i is the perturbation term. The measurement covariance matrix can be estimated as

$$V_k = \frac{\varepsilon_k \cdot \varepsilon_k{}^T}{N-1} \tag{15}$$

Specify to the situation we are dealing with, $A_k = [\,S_k\ E_k\ I_k\ R_k\ \beta_k]^T$ where $S_k = [s_k^1\ s_k^2\ \ldots\ s_k^N]$ and it is the same with E_k, I_k, R_k, and β_k. Note that s_k^i denotes the i^{th} sample in step k. The measurement is the total number of confirmed cases reported by the National Health Commission of China, which is compared with the number of removed population in our model. Therefore, $y_k = C \cdot [x\ \beta]^T$ where $C = [0\ 0\ 0\ 1\ 0]$.

As the criteria for counting the confirmed cases are changed on February 12$^{\text{th}}$, 2020, there is a jump in the reported number of confirmed cases. Taking this into account, we separate the whole time period into period 1 (from January 26$^{\text{th}}$ to February 11$^{\text{th}}$) and period 2 (from February 12$^{\text{th}}$ to February 27$^{\text{th}}$), and run simulations in two steps. First, we run simulations for period 1 and estimate the infection rate β. Then, we take the estimated β on February 11$^{\text{th}}$ and the reported number of cases on February 12$^{\text{th}}$ as the initial β and initial removed population for period 2, respectively, and run simulations. The number of iterations is 250 in every predictive simulation.

2.2. Long-term analysis with the Markovian process

In this section, we apply the generalized modeling framework (GEMF) proposed by Sahneh, Scoglio, & Van Mieghem (2013) to build an individual-level network model. In GEMF, each node stays in a different state, and the joint state of all nodes follow a Markovian process. The node-level description of this Markovian process can be expressed as follows:

$$P_r[x_i(t+\Delta t) = 1 | x_i(t) = 0,\ X(t)] = \beta'(t) Y_i \Delta t \tag{16}$$

$$P_r[x_i(t+\Delta t) = 2 | x_i(t) = 1, X(t)] = \lambda \Delta t + o(\Delta t) \tag{17}$$

$$P_r[X_i(t+\Delta t) = 3 | x_i(t) = 2, X(t)] = \delta \Delta t + o(\Delta t) \tag{18}$$

where $X(t)$ is the joint state of all individual nodes at time t; $x_i(t) = 0,\ 1,\ 2,\ 3$ means that the state of node i at time t is in susceptible, exposed, infectious and removed compartment, respectively; Y_i is the number of infectious neighbors of node i; $\beta'(t)$ is the rate for one node transitions from susceptible state to the exposed state; λ is the rate from the exposed state to the infectious state, and δ is the removal rate from infectious state to the removed state, which is the inverse of the infectious period.

In the network, nodes represent people and are divided to 17 groups to represent the 17 prefecture-level cities in Hubei Province (each group is called a city hereafter). Links are generated to reflect both the effects of person-to person contacts within a city and population flows between the cities on the spread of COVID-19. More specifically, we generate intra-city links for each city in two steps. First, we further divide nodes in each city into multiple communities to make it more realistic. Inside each community, links are generated based on the degree distribution given in Zhang et al. (2020) using the configuration model. Second, on average each node is connected to another node in the same city based on Erdős–Rényi model such that these communities have connections with each other.

Then, we generate inter-city links based on the number of people that traveled between the cities obtained from Baidu Migration data. For example, if the daily population flow from city i to city j is K number of people, directed links will be generated between K randomly selected nodes in city i to Kd_t nodes in city j, where d_t is the network average node degree at time t. More specifically, the average daily population movement from January 1$^{\text{st}}$–22$^{\text{nd}}$ is used to build *inter-city links before the Wuhan City lockdown* on January 23$^{\text{rd}}$. Since the value of population flows between cities remain stable during the quarantine period, we use the average daily population movement from January 23$^{\text{rd}}$–29$^{\text{th}}$ to build *inter-city links after the Wuhan City lockdown*. Details about data collection can be found in Appendix.

In GEMF simulations, we consider the COVID-19 epidemic outbreak starting in Wuhan City from January 19$^{\text{th}}$, because all confirmed cases reported in Hubei Province were in Wuhan City before January 19$^{\text{th}}$, 2020. To reduce the computational cost, the population of each city and population flows among the cities are scaled by 198, which is the number of cumulative confirmed cases in Hubei Province up to January 19$^{\text{th}}$. After scaling, the total number of nodes in the network N equals 298,838. In total, we generate three adjacency lists of the network topology, which are imported to GEMF and changed according to time during simulation. More specifically, the first adjacency list A_1 corresponds to the network with $d_t = 15$ and

inter-city links before the Wuhan City lockdown, and is used during simulation up to January 22nd. The second adjacency list A_2 refers to the network with $d_t = 15$ and inter-city links after the Wuhan City lockdown, and is used during simulation up to February 9th. The third adjacency list A_3 incorporates $d_t = 3$ and inter-city links after the Wuhan City lockdown, and is used during simulation on and after February 10th when community closure was implemented in Wuhan city.

To fit the model, we take the number of confirmed cases as the number of people in the removed state. By keeping $\lambda = 1/3$ day^{-1} and $\delta = 1/4$ day^{-1} unchanged, we estimate the infection rate β' as a piecewise constant function of the time. The β' function and time intervals are both tuned to minimize the error between the number of removed people from our simulations and the number of cumulative confirmed cases in Hubei Province (Health Commission of Hubei Province, 2020). The total period from January 19th to April 15th is separated into three phases (January 19th–January 27th, January 28th–February 9th, and after February 9th). During the first phase, Wuhan city lockdown happened, and hospitals were short of bed. During the second phase, Thunder God Mountain Hospital and Fire God Mountain Hospital were put into use, and door-to-door screening was implemented. In the third phase, community closure was enacted.

GEMF is available in MATLAB, R, Python, and C programming language (Sahneh, Vajdi, Shakeri, Fan, & Scoglio, 2017; Yang, Lu, Scoglio, de Jong, & Gruenbacher, 2018; Yi, Yang, & Scoglio, 2020). We performed extensive stochastic simulations by adapting the GEMF toolbox with MATLAB during the experiment. We track the number of nodes in each compartment and demonstrate the simulation results with confidence intervals. The number of simulation runs is 1500 in all scenarios.

2.3. Long-term analysis with the non-Markovian process

Since the outbreak of COVID-19, many studies have emerged to learn how this disease spreads. However, an overwhelming majority of the epidemic models have considered Markov processes, in which the transition between compartments follows an exponential distribution. To the best of the authors' knowledge, very few works have been done incorporating non-Markovian processes to epidemic models with respect to COVID-19. In this section, we take Wuhan City as a study area and try to demonstrate how distributions of key epidemiological factors can influence the predicted trajectory of the disease spreading.

More specifically, we incorporate non-exponential interevent time distributions to the SEIR model using the modified Gillespie algorithm proposed by Boguñá, Lafuerza, Toral, & Serrano (2014). Transitions from the susceptible state to the exposed state is a Poisson process that follows the exponential distribution. Following lognormal distributions, the mean and median of the incubation period are estimated at 5.6 days and 5.0 days, respectively, while the infectious period is of mean 3.9 days and median 1.5 days (Linton et al., 2020), which is in line with Sanche et al. (2020). The number of simulation runs is 80.

3. Results and discussion

In this section, we provide results regarding the short-term analysis with the Kalman filter. From the long-term perspective, we first present results based on GEMF stochastic simulations and then compare the results based on non-Markovian processes and Markov processes.

3.1. Short-term analysis with the Kalman filter

The predictions of COVID-19 cases in Wuhan city from February 22nd to March 1st are conducted. The transition rates λ and δ are 1/3 day^{-1} and 1/4 day^{-1}, respectively. The infection rate is estimated by EnKF. The model noise and measurement noise are tuned by modifying the approach proposed by Li, Pei, et al. (2020).

Fig. 1 shows the distributions of the predicted numbers of COVID-19 cases for 1-day-ahead, 2-day-ahead, and 3-day-ahead. The solid black lines indicate the maximum and minimum prediction results on each day. The blue box presents the range from the lower quartile to the upper one. The red line inside the box indicates the median of the samples. We can see that most predictions converge to a small range, and that most medians are very close to the corresponding real data. Taking the median as our predicted number gives the relative error shown in Fig. 1 (d). The mean absolute percentage error (MAPE) for 1-day-ahead prediction is 0.24%; the MAPE for 2-day-ahead prediction is 0.52%; the one for 3-day-ahead is 1.23%. The performance of 1-day-ahead prediction is the best and the one of 3-day-ahead is the worst. When the n-day-ahead prediction is conducted, the results from (n-1)-day-ahead is used, so the errors accumulate.

If we choose the average number of the samples as our predicted number of COVID-19 cases, the results and their relative errors are presented in Fig. 2. The MAPE for 1, 2, 3-day-ahead predictions are 0.24%, 0.52% and 1.23%, which are equal to the ones when the medians are predicted numbers.

The self-quarantine policy was carried out on January 23rd in Wuhan and a series of administrative actions were taken to prevent the disease from spreading. Taking the latent period into consideration, we started to estimate the infection rate β from January 26th. On February 12th, Wuhan authorities changed the case definition to include symptomatic patients in addition to tested ones. Therefore, we present estimations of β for period 1 (from January 26th to February 11th) and period 2 (from February 12th to February 27th), which are presented as the blue line and the red line in Fig. 3. The orange line in Fig. 3 presents the removal rate. From the results, we see that the infection rate β is less than the removal rate δ after February 10th, which means the epidemic had peaked in Wuhan.

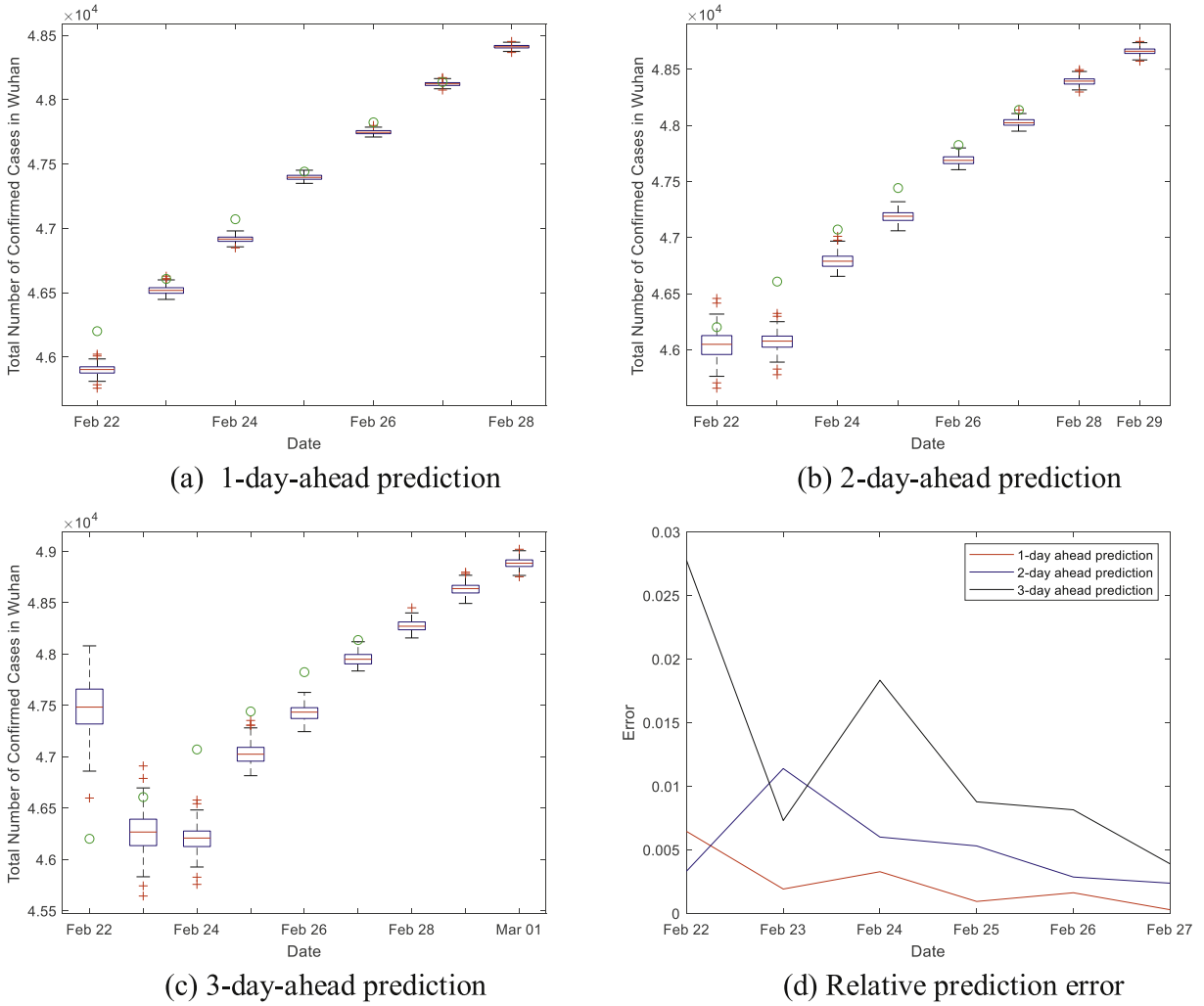


Fig. 1. Distributions of the short-term prediction.

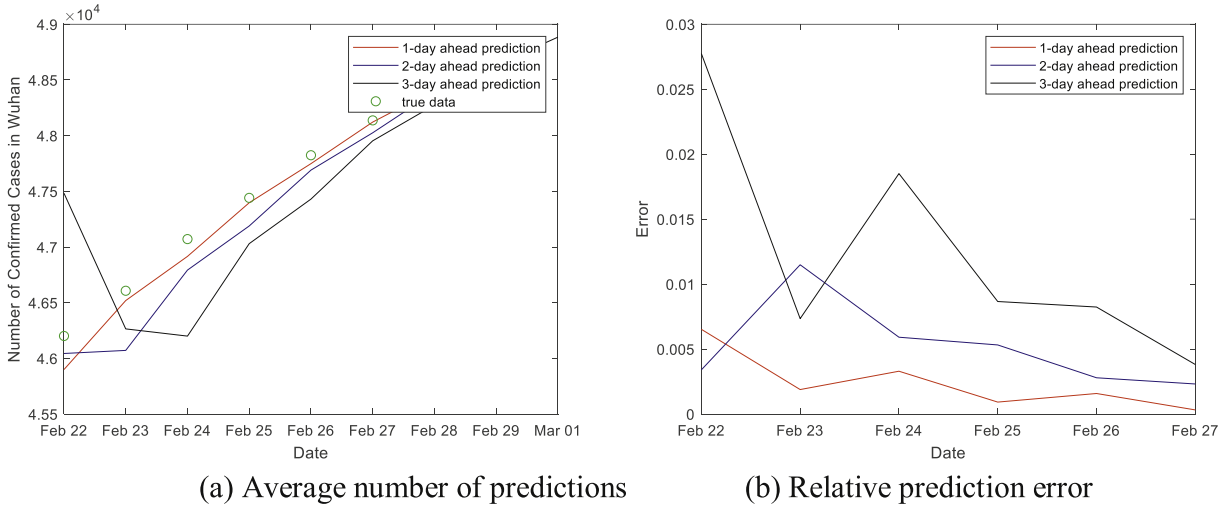


Fig. 2. Average number as predictions and the error.

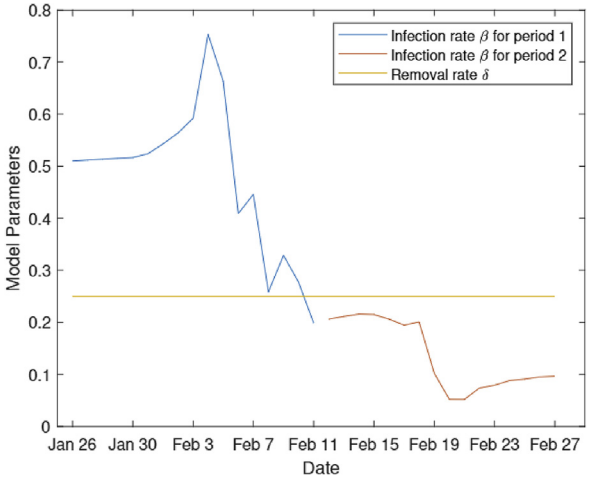


Fig. 3. Estimated infection rate.

3.2. Long-term analysis with the Markovian process

To reconstruct the epidemics in Hubei province, we continuously tune the infection rate parameter $\beta^{'}$ to minimize the error between the number of nodes in the removed compartment from GEMF simulations and the number of cumulative confirmed cases from the data, as shown in Fig. 4.

By performing GEMF stochastic simulations, we obtain that $\beta^{'}$ is equal to 0.06 for day 0–8 (January 19th–January 27th), and then becomes 0.0169 for day 9–21 (January 28th–February 9th), and equals 0.01 on and after day 22 (February 10th). The infection rate dramatically drops, indicating that public health interventions implemented in Hubei Province have considerably controlled the epidemic development since the nascent stage of the outbreak. Based on the fitted parameter $\beta^{'}$, we estimate the basic reproduction number ($R_0 = \beta^{'} d_t/\delta$) decreases from 3.6 before January 27th, to 1.014 between January 27th–February 9th, and to 0.12 on and after February 10th.

Fig. 5 shows the epidemic trajectories under four scenarios, in which different mitigation strategies are implemented after February 9th: continuing the current trend with $\beta^{'} = 0.0169$ and the network average node degree $d_t = 15$ (Scenario 1); reducing d_t from 15 to 3 (Scenario 2); decreasing $\beta^{'}$ by 75% (Scenario 3); reducing $\beta^{'}$ from 0.0169 to 0.01 and d_t from 15 to 3 (Scenario 4). Preventive measures such as wearing masks can help reduce the infection rate. On the other hand, the average node degree refers to the possible contacts a person may have in daily life, and this can be reduced by contact reduction and social distancing.

In Fig. 5, if the current trend as of February 9th had continued, the epidemic spreading could have become more severe than what happened in Hubei Province. Reducing the average node degree from 15 to 3 (Scenario 2), the epidemic would reach the peak at around day 21.85 and fade out at around day 66.80 with the final epidemic size 367.86 (95% 339.55–396.18).

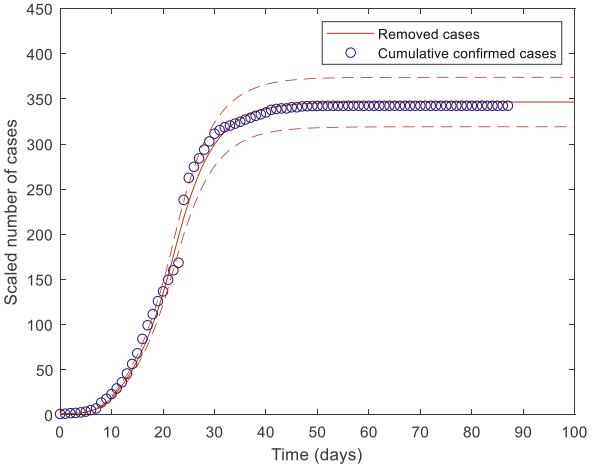
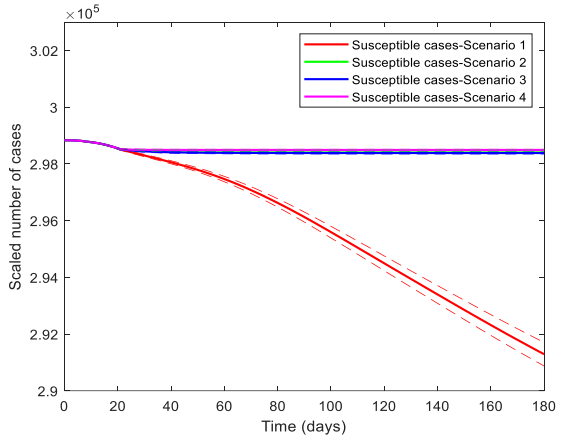
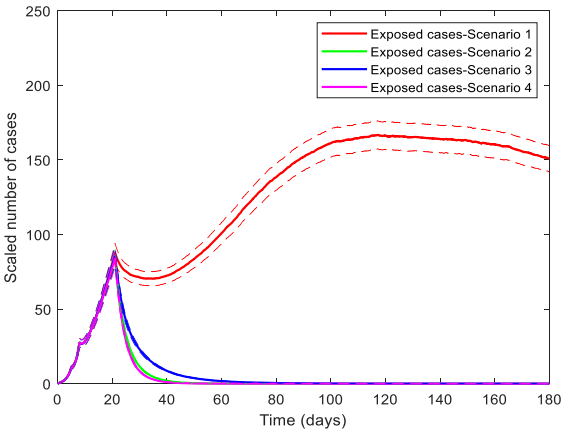


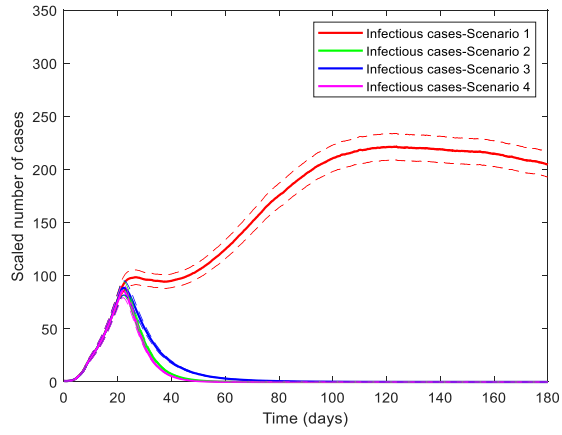
Fig. 4. Fitting the model to the cumulative confirmed cases from January 19th to April 15th.



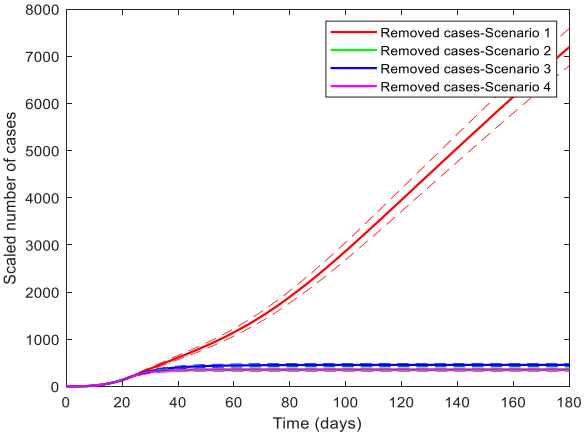
(a) Scaled number of susceptible cases



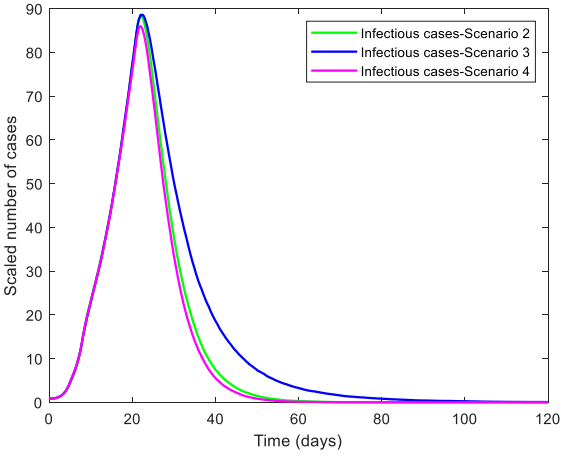
(b) Scaled number of exposed cases



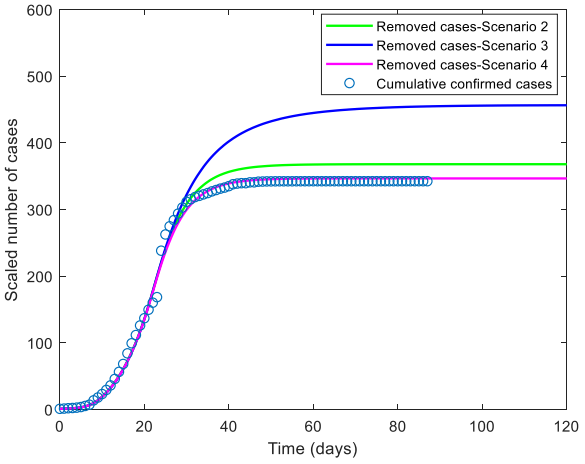
(c) Scaled number of infectious cases



(d) Scaled number of removed cases



(e) Scaled average number of infectious cases after zooming in



(f) Scaled average number of removed cases compared to the cumulative confirmed case

Fig. 5. Epidemic dynamics under four scenarios (red, green, blue, and magenta) from GEMF stochastic simulations. Note that the magenta curves almost overlap perfectly with the green curves in Fig. 5(c), so Fig. 5(e and f) are provided to zoom in and show the differences.

When $\beta^{'}$ is reduced by 75% in Scenario 3, the number of removed cases in the steady-state would reach 456.47 on average (95% CI: 423.34–489.59), and the epidemic would fade out at day 110.90 on average. When a combination of reducing $\beta^{'}$ and average node degree is conducted in Scenario 4, the epidemic dynamics would approximate the actual epidemic trajectory in Hubei Province in the steady-state with mean scaled number of removed cases 346.54 (95% CI: 319.12–373.96). In addition, the epidemic would reach the peak at day 21.50 and fade out at day 61.55 on average.

Overall, GEMF simulation results suggest that more strict prevention and control measures have been taken since February 9th to achieve the current cumulative confirmed cases in Hubei Province. Preventive measures and social distancing measures need to be taken in combination by the government to speed up the containment of the epidemic and reduce the peak infection dramatically.

3.3. Long-term analysis with the non-Markovian process

With mean infectious period (from compartments $I \rightarrow R$) set as 3.9 days and mean incubation period (from compartments $E \rightarrow I$) set as 5 days, we first use ordinary differential equations based on Markov processes to estimate the infection rate β from compartments $S \rightarrow E$. Assuming that the epidemic starts from December 1st, β is equal to 0.68 to match the number of cumulative confirmed cases in Wuhan city until February 12th.

Second, we simulate the disease spread in Wuhan city with transmission between compartments incorporating non-Markovian dynamics. The transition from compartment $S \rightarrow E$ still evolves according to an exponential distribution with

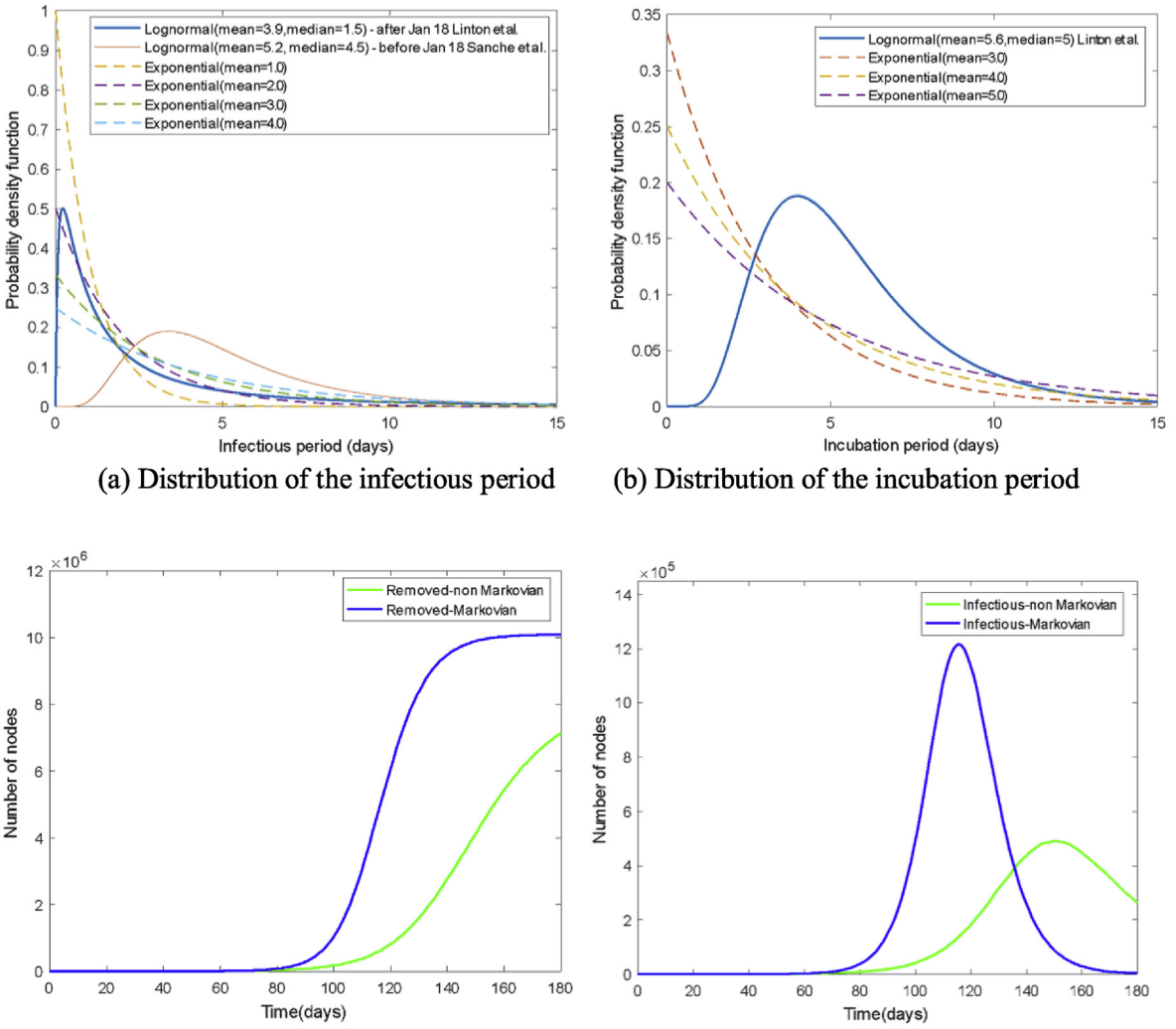


Fig. 6. Comparison between results based Markovian process and non-Markovian process.

$\beta = 0.68$, while transitions from compartments $E \rightarrow I$ and $I \rightarrow R$ follow the lognormal distributions from (Linton et al., 2020; Sanche et al., 2020) in Fig. 6 (a and b). Note that the width of the infectious period becomes much shorter (blue curve) after January 18th, as shown in Fig. 6 (a), because the government has started to trace back and monitor those people who have contact with confirmed cases and built temporary hospitals to hospitalize both patients with mild and severe symptoms since mid-January. In Fig. 6 (a and b), we also show curves of different exponential distributions compared with these lognormal distributions for demonstration purposes.

In Fig. 6 (c and d), simulated results incorporating non-Markovian processes deviate from the results obtained from Markov processes. More specifically, the epidemic curve with non-Markovian processes is flattened with a smaller epidemic peak compared with that from Markov processes. This is because the infectious period used in the Markovian model has an exponential distribution (mean = 3.9 days) similar to the shape of the exponential distribution (mean = 4.0 days) in Fig. 6 (a), but shorter than the lognormal distribution (mean = 3.9 days, median = 1.5 days) used after January 18th. The reasons above indicate that control measures to reduce the infectious period can effectively mitigate the outbreak.

On the other hand, these two methods with different distributions but with the same mean infectious period lead to inconsistent results. Since the actual infectious period and incubation period do not follow exponential distributions, these simulations suggest that future works need to consider distributions of epidemiological characteristics to better capture the COVID-19 spreading trajectory. Instead of taking mean values derived from clinical studies as input to the epidemic models, researchers need to take into account the empirically derived distribution of the parameters.

4. Conclusion

In this work, the ensemble Kalman filter is used to make short-term predictions of the COVID-19 cases in Wuhan City, which resulted in accurate forecasts of daily case reports. The model is able to predict daily cases and the epidemic peak. Knowing the daily cases from forecasts three days in advance allows for the proper resource allocation. The ensemble Kalman filter also allows parameter estimation, which is extremely useful for modeling purposes.

From an epidemiological modeling perspective, we simulate the epidemic spreading in Wuhan city based on non-Markovian processes, and show that different distributions with the same mean infectious period can lead to inconsistency in terms of epidemic dynamics. Since most current network-based approaches are based on Markov processes, in which epidemiological parameters are assumed to follow exponential distributions, non-Markovian process-based models need to be further explored to better predict the COVID-19 spreading. Our results show that reducing the infectious period via measures such as early case identification and isolation can reduce the epidemic size significantly.

In the long-term perspective, we test and confirm the effectiveness of non-pharmaceutical interventions on the containment of the epidemic dynamics by performing GEMF stochastic simulations. The daily cases of the COVID-19 epidemic in Hubei Province peaked at over 15,000 on February 12th and became almost zero in mid-March (Xinhua, 2020). Our simulation results indicate that the containment of this highly contagious disease was achieved by stringent control measures by the Chinese government. Without continued and aggressive control measures, the epidemic in Hubei Province would have become more severe. Only with combined implementation of enhanced protective measures and social distancing measures, the epidemic dynamics would peak at around mid-February and approximate the actual epidemic trajectory in March. This can be an important message for countries going through the exponential growth of the epidemic in the current days.

Declaration of competing interest

None.

Acknowledgement

This work was supported by the Department of the Army, U.S. Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Ft Detrick, MD (DWFP grant W911QY-19-1-0004) and the National Science Foundation under Grant Award IIS-2027336. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

Appendix

A. Method

A1. Data collection

The population size of each city in Hubei Province is obtained from the Hubei Statistical Yearbook (Hubei Provincial Bureau of Statistics, 2020). We collected Baidu Migration data (https://qianxi.baidu.com/) for the 17 cities in the Hubei Province from January 1st to January 29th, 2020. The daily Baidu migration data contain two types of data. First type: the data include the immigration index and emigration index of each city, which are linearly related to the human traffic volume moving in and

out of individual cities. Second type: for each city i, the data also contain the composition of city i's incoming flow, namely the fraction of flows coming from every other city j to city i (flow from city j to i divided by the total flow coming into city i). Similarly, the data contain the fraction of flows going out of city i to city m (flow from city i to m divided by the total flow going out of city i).

Based on the fact that around 5 million people have moved out of Wuhan before the city lockdown on January 23rd (Zhu et al., 2020), we obtain the number of people that traveled between the cities using the following approach. First, we estimate value of α, which equals 5 million divided by the sum of Wuhan's emigration indexes from January 1st to January 23rd. Then, we can calculate the population flow moving in or out of each individual city by α times the immigration or emigration index. In addition, the population flow multiplied by the flow fraction going to or coming from a specific city gives us the number of daily travelers to or from a particular individual city to another.

References

Boguñá, M., Lafuerza, L. F., Toral, R., & Serrano, M.Á. (2014). Simulating non-Markovian stochastic processes. *Physical Review A, 90*(4). https://doi.org/10.1103/PhysRevE.90.042108.

Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *The Lancet, 395*(10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, China: A descriptive study. *The Lancet, 395*(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., et al. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*. https://doi.org/10.1126/science.aba9757.

Du, Z., Wang, L., Cauchemez, S., Xu, X., Wang, X., Cowling, B. J., et al. (2020). Risk for transportation of 2019 novel coronavirus disease from wuhan to other cities in China. *Emerging Infectious Diseases Journal, 26*(5), 1049–1052. https://dx.doi.org/10.3201/eid2605.200146.

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2002032.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet, 395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

Hubei Provincial Bureau of Statistics. (2020). Hubei province statistical Yearbook 2019. Retrieved March 2, 2020, from http://tjj.hubei.gov.cn/tjsj/sjkscx/tjnj/qstjnj/.

http://wjw.hubei.gov.cn/, 2020. (Accessed 2 March 2020).

http://www.xinhuanet.com/english/2020-03/13/c_138871417.htm, 2020. (Accessed 26 March 2020).

Jazwinski, A. H. (2007). *Stochastic processes and filtering theory*. Courier Corporation.

Jensen, J. P. (2007). *Ensemble kalman filtering for state and parameter estimation on a reservoir model*. Master thesis. Retrieved from http://hdl.handle.net/11250/259624.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45. https://doi.org/10.1115/1.3662552.

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., et al. (2020). Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak in China. *MedRxiv*. https://doi.org/10.1101/2020.03.03.20029843.

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S., et al. (2020). incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine, 9*(2), 538. https://doi.org/10.3390/jcm9020538.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., et al. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science, 368*(6490), 489–493. https://doi.org/10.1126/science.abb3221.

Li, X., Zhao, X., & Sun, Y. (2020). *The lockdown of Hubei Province causing different transmission dynamics of the novel coronavirus (2019-nCoV) in Wuhan and Beijing*. MedRxiv. https://doi.org/10.1101/2020.02.09.20021477.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., et al. (2020). Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa2001316.

Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S. S., et al. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases, 93*, 211–216. https://doi.org/10.1016/j.ijid.2020.02.058.

Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state–parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources, 28*(2), 135–147. https://doi.org/10.1016/j.advwatres.2004.09.002.

Nishiura, H., Jung, S., Linton, N. M., Kinoshita, R., Yang, Y., Hayashi, K., et al. (2020). The extent of transmission of novel coronavirus in wuhan, China, 2020. *Journal of Clinical Medicine, 9*(2), 330. https://doi.org/10.3390/jcm9020330.

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., & Jewell, C. P. (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*. https://doi.org/10.1101/2020.01.23.20018549.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., et al. (2020). Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New England Journal of Medicine, 382*(10), 970–971. https://doi.org/10.1056/NEJMc2001468.

Sahneh, F. D., Scoglio, C., & Van Mieghem, P. (2013). Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking, 21*(5), 1609–1620. https://doi.org/10.1109/TNET.2013.2239658.

Sahneh, F. D., Vajdi, A., Shakeri, H., Fan, F., & Scoglio, C. (2017). GEMFsim: A stochastic simulator for the generalized epidemic modeling framework. *Journal of Computational Science, 22*, 36–44. https://doi.org/10.1016/j.jocs.2017.08.014.

Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N. W., & Ke, R. (2020). The novel coronavirus, 2019-nCoV, is highly contagious and more infectious than initially estimated. *ArXiv*. https://arxiv.org/abs/2002.03268.

Shen, M., Peng, Z., Guo, Y., Xiao, Y., & Zhang, L. (2020). Lockdown may partially halt the spread of 2019 novel coronavirus in Hubei province, China. *MedRxiv*. https://doi.org/10.1101/2020.02.11.20022236.

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, China: A modelling study. *The Lancet*. https://doi.org/10.1016/S0140-6736(20)30260-9.

Wu, T., Ge, X., Yu, G., & Hu, E. (2020). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *MedRxiv*. https://doi.org/10.1101/2020.02.25.20027433.

Yang, Q., Lu, H., Scoglio, C., de Jong, M., & Gruenbacher, D. (2018). A network-based approach to model the development of city branding in China. *Computers, Environment and Urban Systems, 72*, 161–176. https://doi.org/10.1016/j.compenvurbsys.2018.06.007.

Yi, C., Yang, Q., & Scoglio, C. M. (2020). Understanding the effects of the direct contacts and the indirect contacts on the epidemic spreading among beef cattle farms in southwest Kansas. *BioRxiv*. https://doi.org/10.1101/2020.04.27.063735.

Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., … Yu, H. (2020). Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science, 368*(6498). https://doi.org/10.1126/science.abb8001.

Zhao, S., Lin, Q., Ran, J., Musa, S. S., Yang, G., Wang, W., et al. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases, 92*, 214–217. https://doi.org/10.1016/j.ijid.2020.01.050.

Zhu, X., Zhang, A., Xu, S., Jia, P., Tan, X., Tian, J., et al. (2020). Spatially explicit modeling of 2019-nCoV epidemic trend based on mobile phone data in mainland China. *MedRxiv*. https://doi.org/10.1101/2020.02.09.20021360.

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.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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	$2.5° \times 2.5°$
Avg surface relative humidity [c]	Daily	$2.5° \times 2.5°$
Avg surface pressure [c]	Daily	$2.5° \times 2.5°$
Avg precipitable water [c]	Daily	$2.5° \times 2.5°$

[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).

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

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

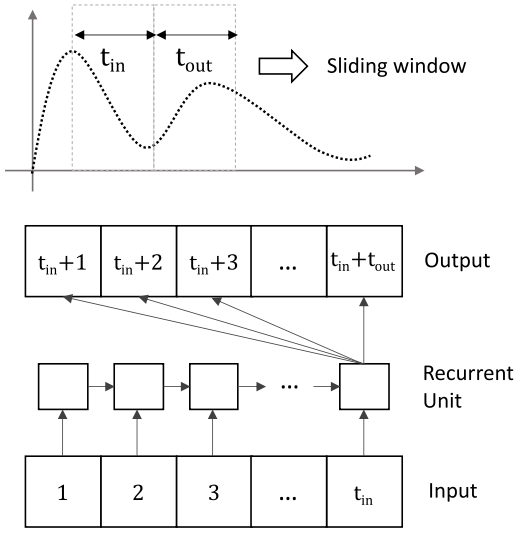


FIGURE 1. A sliding window defined for feeding of input data and extraction of output case counts. The window slides along the horizontal axis and feeds t_{in} time steps of feature data into the model and extracts t_{out} time steps of predictions.

neural network model as a trivial baseline. Using the TensorFlow [46] package, we configure the models as described below to produce the results:

- *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

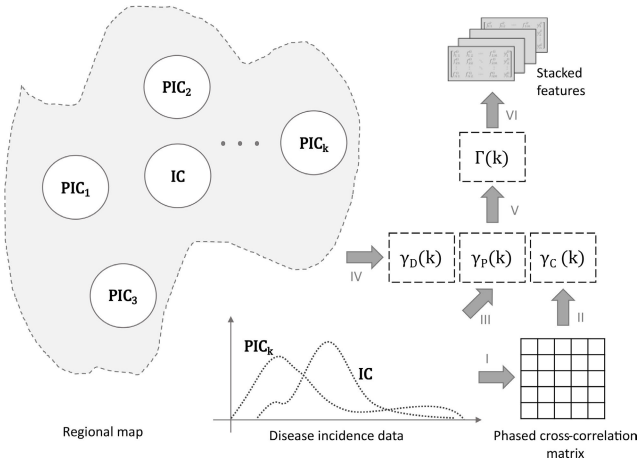


FIGURE 2. A framework depicting the method to expand trainable and testable data set on dengue incidence. The shaded region enclosed by dashed borders on the left depicts a map of the region of interest. The circles inside the map indicate multiple infection centers (*IC*) across the region. The target 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 *ICs*, 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

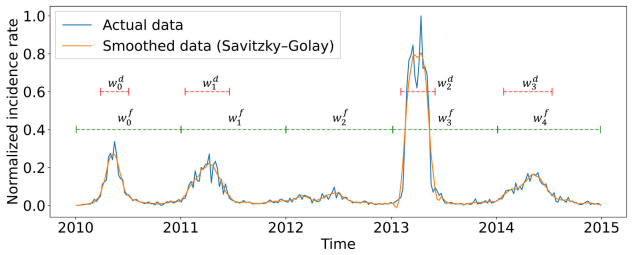


FIGURE 3. Windowing methods used for data segmentation before computing time-shifted correlation coefficients. We present the normalized incidence data from the *Cachoeiro de Itapemirim* municipality of Brazil [42]. The fixed windows (w_m^f) are marked in green, and the detected windows (w_m^d) are marked 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.

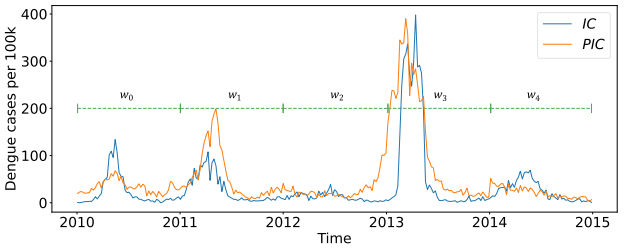


FIGURE 4. Windowing of *IC* and *PIC* incidence data for computing time-shifted cross-correlation coefficients. We allocate a fixed number of windows ($M^f = M = 5$) for the time range 2010-2015, making each window 1 year long (52 weeks approximately). The two curves shown here correspond to the target *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*.

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

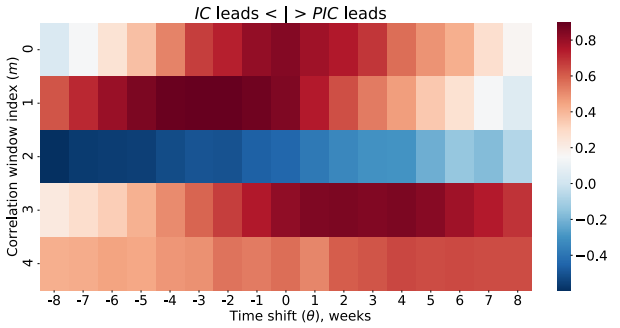


FIGURE 5. Heatmap of the computed time-shifted cross-correlation matrix (3) for the *IC* and the *PIC* in Figure 4. The vertical axis depicts the window indices ($m \in M$) and the horizontal axis depicts time shift (phase), θ that ranges from -8 to $+8$ weeks. The color shades of the heatmap depict 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*).

window (w_m), we find the location (θ) of the peak correlation as shown in (4).

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

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 *PIC*s. 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 *PIC*s 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 *PIC*s based on Γ are listed in Table 3 with the corresponding weights. While our method prioritizes the correlation weight more than others, *PIC*s 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 *PIC*s 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 *PIC*s are mostly correlated with the *IC*, Vitória. Among the *PIC*s 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

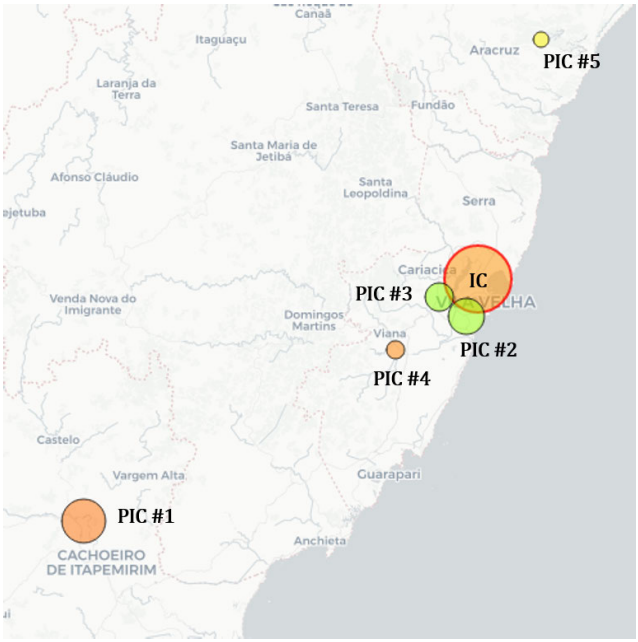


FIGURE 6. A geospatial map showing an incidence center (*IC*) and 5 ranked peripheral incidence centers (*PIC*). The circles' radii are proportional to the total number of cases reported during 2010-2019 [42]. The shades of the fill color are generated from a color gradient (green-yellow-red) which is 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].

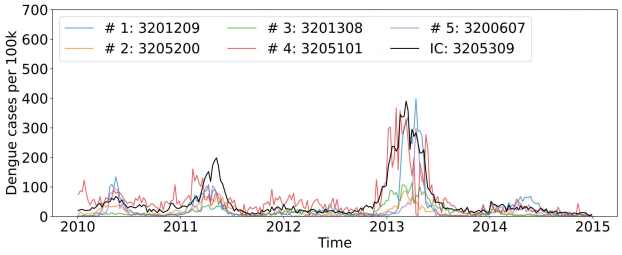


FIGURE 7. Time series plots of weekly dengue cases per 100,000 people for the *IC* and top 5 ranked *PIC*s (Table 3). The plots depict cases only between 2010-2015 for improved clarity, but the metrics were computed based on the entire series (2010-2019).

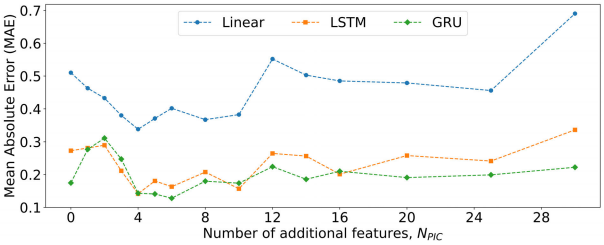


FIGURE 8. Prediction performances of the Linear, LSTM, and GRU models in predicting normalized test data for varying number of features (N_{PIC}). *PIC* data are added to the feature set based on ranks dictated by computed predictor strengths (Table 3), after which models are trained and evaluated over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

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.

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

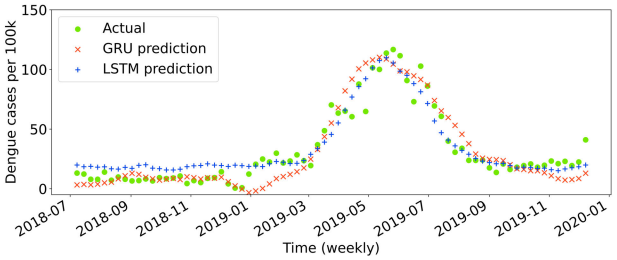


FIGURE 9. Predicted weekly dengue cases per 100,000 people using test data with the recurrent models (LSTM and GRU). These results were obtained for the *IC*, Vitória, with 4 and 6 additional *PIC* data joined with LSTM and GRU's feature set, respectively.

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 *IC*s based on total cases per 100,000 people and average the performance metrics across *IC*s. Among the 20 *IC*s, there were 4 *IC*s 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 *IC*s) 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.

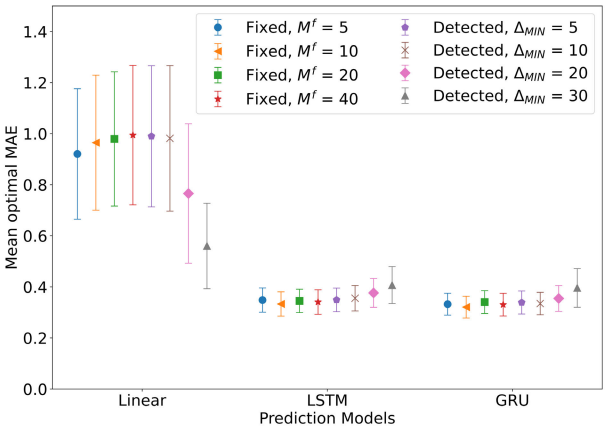


FIGURE 10. A comparison chart of the lowest prediction error obtainable using different fixed and detected windowing schemes. Four fixed windowing schemes with the number of windows, $M^f = 5$, 10, 20, and 40, along with four window detection schemes having minimum window lengths, $\Delta_{MIN} = 5$, 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 *IC*s 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 *IC*s, however, a conclusive determination of the best performing model can be done (either GRU or LSTM).

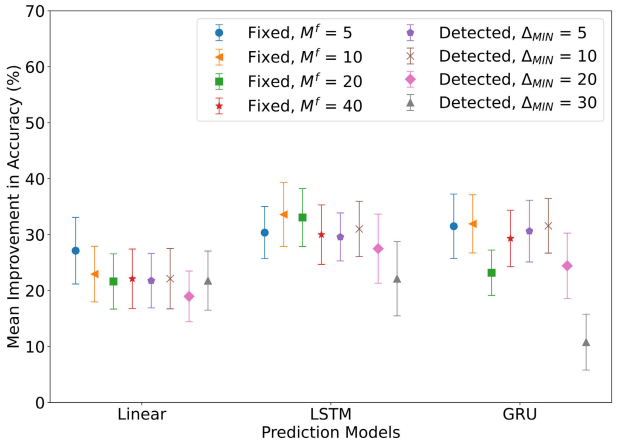


FIGURE 11. A comparison chart of average improvement in accuracy (reduction in MAE) with respect to performance without feature enhancements ($N_{PIC} = 0$) for different fixed and detected windowing schemes. Four fixed windowing schemes with the number of windows, $M^f = 5$, 10, 20, and 40, along with 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.

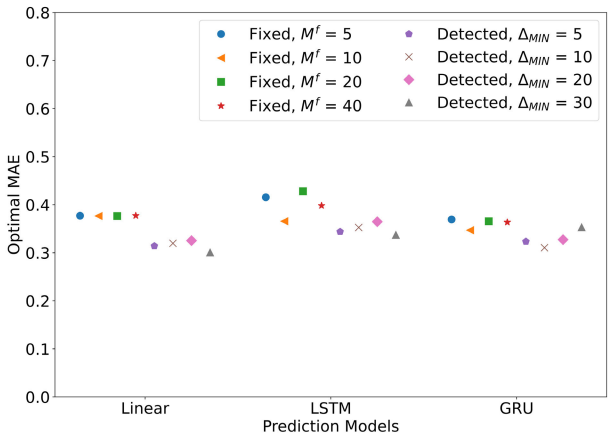


FIGURE 12. A comparison chart of lowest prediction error obtainable using different fixed and detected windowing schemes for case data aggregated across an ecoregion. Four fixed windowing schemes with the number of windows, $M^f = 5$, 10, 20, and 40, along with four window detection schemes having minimum 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

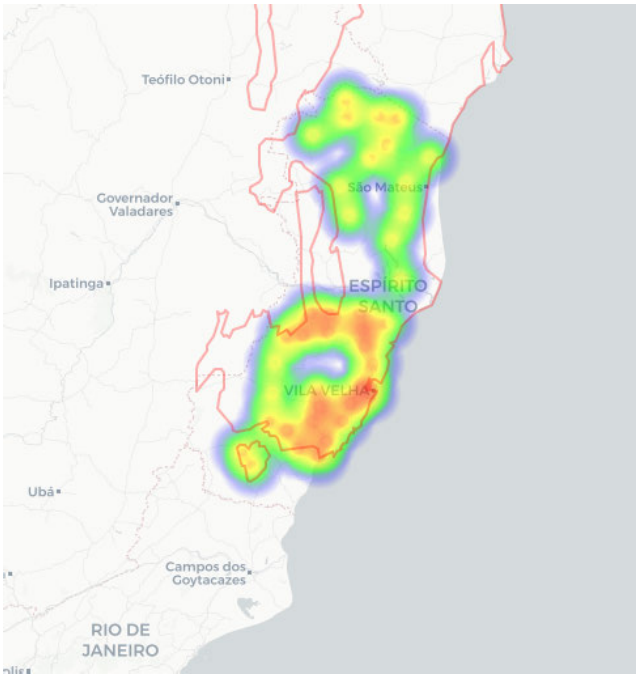


FIGURE 13. A geospatial map showing the predicted risk of infection in an ecoregion of Brazil. A part of the ecoregion (Bahia Coastal Forests) shown here is marked with red borders. The risk of infection is shown as a heatmap with color shades ranging from blue (low risk) to red (high risk). The heatmap 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 (*PIC*s) 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 *IC*s, 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 *PIC*s are included in the map, and the *PIC*s 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.

pathogens

MDPI

Article

PICTUREE—Aedes: A Web Application for Dengue Data Visualization and Case Prediction

Chunlin Yi [1], Aram Vajdi [1], Tanvir Ferdousi [1,†], Lee W. Cohnstaedt [2,*] and Caterina Scoglio [1]

[1] Department of Electrical and Computer Engineering, College of Engineering, Kansas State University, Manhattan, KS 66506, USA; cyi@ksu.edu (C.Y.); avajdi@ksu.edu (A.V.); tanvirf@ksu.edu (T.F.); caterina@ksu.edu (C.S.)
[2] National Bio- and Agro-Defense Facility, Agricultural Research Service, United States Department of Agriculture, Manhattan, KS 66502, USA
* Correspondence: lee.cohnstaedt@usda.gov
† Current address: Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22911, USA.

Abstract: Dengue fever remains a significant public health concern in many tropical and subtropical countries, and there is still a need for a system that can effectively combine global risk assessment with timely incidence forecasting. This research describes an integrated application called PICTUREE—Aedes, which can collect and analyze dengue-related data, display simulation results, and forecast outbreak incidence. PICTUREE—Aedes automatically updates global temperature and precipitation data and contains historical records of dengue incidence (1960–2012) and *Aedes* mosquito occurrences (1960–2014) in its database. The application utilizes a mosquito population model to estimate mosquito abundance, dengue reproduction number, and dengue risk. To predict future dengue outbreak incidence, PICTUREE—Aedes applies various forecasting techniques, including the ensemble Kalman filter, recurrent neural network, particle filter, and super ensemble forecast, which are all based on user-entered case data. The PICTUREE—Aedes' risk estimation identifies favorable conditions for potential dengue outbreaks, and its forecasting accuracy is validated by available outbreak data from Cambodia.

Keywords: web application; dengue transmission; mosquito population estimation; dengue risk estimation; dengue incidence forecasts

check for updates

Citation: Yi, C.; Vajdi, A.; Ferdousi, T.; Cohnstaedt, L.W.; Scoglio, C. PICTUREE—Aedes: A Web Application for Dengue Data Visualization and Case Prediction. *Pathogens* **2023**, *12*, 771. https://doi.org/10.3390/pathogens12060771

Academic Editor: Stephanie L. Richards

Received: 21 April 2023
Revised: 18 May 2023
Accepted: 19 May 2023
Published: 29 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The PICTUREE—Aedes web application is an innovative system that integrates tools for displaying dengue-related data, visualizing simulated results, and forecasting future cases for an ongoing outbreak. This application is a product of the Predicting Insect Contact and Transmission Using histoRical Entomological and Environmental data (PICTUREE) project, which was initiated in 2019.

Dengue fever is an *Aedes* mosquito-borne viral infection that threatens the health of over 3 billion people in numerous tropical and subtropical regions [1]. *Aedes aegypti* mosquito is the primary carrier of dengue viruses, with the ability to bite multiple individuals in a brief period, and its immature phase can be found in water containers near human dwellings [2]. The proliferation, biting rates, infective rate, and survivability rate of *Aedes* mosquitoes are substantially affected by climatic conditions such as rainfall, temperature, and relative humidity, which, in turn, have a significant impact on the spread of the dengue virus [3–5]. The breeding time and maturity period of *Aedes* mosquitoes are shortened due to global warming, resulting in the accelerated growth of the vector population [6]. Furthermore, increasing human mobility and high population density have facilitated the geographical spread of dengue viruses over long distances [7]. Therefore, an integrated

system is needed to aggregate collected data, analyze them, estimate parameters, and forecast future risks and incidence to proactively combat the spread of the virus.

Numerous research studies have been conducted to identify and quantify the temperature-dependent variables that impact the life cycle of the mosquito vector, due to the susceptibility of the mosquito's life cycle to temperature [8–11]. Mathematical equations can be used to model the dynamics of the mosquito population and to estimate the parameters such as mortality, transition rate, and oviposition rate at different temperatures [5]. Another dengue transmission model, which took into account various temperatures, survival rates, between-bite intervals, and extrinsic incubation periods, discovered that an increase in mosquito longevity and temperature led to an augmentation in dengue transmission [12].

There has been an increasing number of approaches to predicting the incidence of infectious diseases, including the use of statistical models [13,14], state-space compartmental models [15–18], deep learning methods [19–21], etc. The state-space models coupled with data assimilation algorithms can generate real-time forecasts and estimate key parameters. The estimated model parameters can then be utilized to determine the basic reproduction number of the disease, the carrying capacity of the vector population, and other relevant metrics. A state-space compartmental disease model combined with an ensemble Kalman filter (EnKF) was employed to estimate and predict future dengue incidence, infection rates, and mosquito emergence rates in [16]. Deep learning approaches, especially the long short-term memory (LSTM) recurrent neural network, are also widely used to generate reliable epidemic forecasts [21]. Machine learning approaches have the advantage of being able to handle nonlinearity and complex data structures. Another emerging technique is the ensemble forecast that enhances the forecasting accuracy of individual models [22,23]. A superensemble forecast was developed by combining three models—susceptible, infectious, recovered, ensemble adjustment Kalman filter, Bayesian weighted outbreaks, and historical likelihood—using the weighted average method [23]. The superensemble was used to predict dengue cases in San Juan, and it was found to provide more accurate forecasts compared to individual models. The use of ensemble forecasts, which combine multiple forecasting models or techniques, can help reduce errors and biases that may be present in individual forecasts.

The PICTUREE—Aedes web application aggregates dengue-related data from multiple sources, including global daily temperature, global daily precipitation, *Aedes* mosquito occurrence, and dengue case occurrence. These data are stored in an SQL database and can be readily visualized through user-friendly displays. The PICTUREE—Aedes web application also provides estimations of the *Aedes* mosquito population, the dengue transmission reproduction number R_t, and the dengue risk globally, which can be viewed by users in both space and time. Furthermore, the application employs several algorithms, including the ensemble Kalman filter, the recurrent neural network, the particle filter, and the super ensemble, to forecast dengue cases for an ongoing outbreak. To generate forecasts, users must provide information about the cumulative reported cases in recent weeks, the outbreak location, and the local population size. After processing this data, the application outputs the statistics of the three-week forecasted cases.

2. Data

Table 1 presents a comprehensive list of the data sources. The daily global temperature and precipitation data are obtained from the National Oceanic and Atmospheric Administration (NOAA) weather stations [24]. The ERA5 hourly data provide hourly estimates of atmospheric variables, such as temperature, pressure, and humidity, which are used for simulation purposes [25]. The terrestrial ecoregions data are utilized to select areas with similar climate conditions [26–28]. The *Aedes* mosquito occurrence dataset and dengue occurrence dataset are geo-positioned occurrence records spanning from 1960 to 2012 [29,30]. Country-wise Google trend data for the keywords "dengue" and "fever" are also collected. The aforementioned data are either directly displayed on the map or are used for the simulations and forecasts [31].

Table 1. Data sources.

Item	Source	Remark
Daily weather observation	[24]	Updated weekly
ERA5 hourly data	[25]	2018–present
Terrestrial ecoregions	[26–28]	As of 14 December 2009
Aedes mosquito occurrence	[29]	1960–2014
Dengue occurrence	[30]	1960–2012
Country-wise Google trends	[31]	2015–2022

3. Simulation Tools

The PICTUREE-Aedes simulations consist of two primary components: (1) the estimations of dengue risk-related measures and (2) the forecasts of case number for an ongoing outbreak. The tools graphically represent current and future risk of dengue transmission for users based on the life history of the mosquito vector.

3.1. Estimating Dengue Risk-Related Measures

We developed a dengue transmission risk model by training a neural network classifier over historical dengue outbreak data. The model uses the estimated vector population and the Dengue transmission reproductive number as predictive variables. In order to estimate the vector population and reproductive number, we developed a mechanistic model to simulate the *Ae. aegypti* population. The mechanistic model incorporates the mosquito's temperature-dependent parameters in a system of integral equations to calculate the mosquito population at various stages of development, including immature and adult stages. A diagram of the mechanistic model is shown in Figure 1. The set of integral equations precisely considers the temperature-dependent development time required for mosquitoes to transform from an egg into adult form. Given that the development period tends to be longer in seasons with cooler temperatures, such precise calculation is especially crucial in estimating the mosquito population in areas that experience significant temperature fluctuations throughout the year.

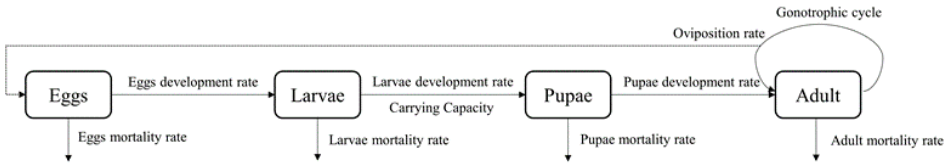


Figure 1. Life cycle of *Aedes aegypti* and the parameters influencing the transition between different stages.

3.1.1. Vector Population

To accurately simulate the mosquito population fluctuations, the model accounts for the time spent in each life history stage during the aquatic or immature portion of the mosquito life cycle, as shown in Figure 1. The development rates of these stages are temperature-dependent, with warmer temperatures resulting in faster development. The temperature-dependent rates were determined from experimental data [5,8,10]. In addition to temperature, the effect of precipitation on the mosquito population is increased through the carrying capacity of the location: more rainfall equates to greater availability of habitats or more containers filled with water. Alternatively, less rainfall in some urban areas results in more irrigation or water storage, leading to an increase in filled containers, but this is location dependent. The equations that describe the life cycle of the mosquitoes are non-Markovian, meaning that transition times to new stages in the cycle are not exponential random variables. The transition to a new stage happens when the development of a current stage is complete. Moreover, the development rates are strongly temperature-dependent,

and temperature changes throughout the year. Therefore, we use integral equations to describe the life cycle of the mosquitoes accurately.

In order to estimate the impact of precipitation on carrying capacity, the life cycle equations are extended to account for virus transmission and to subsequently fit the model to various dengue outbreak data from around the world using particle filter smoothing. This enables us to indirectly estimate the carrying capacity and its variation with precipitation. For this process, we used 67 dengue epidemic curves from Brazil, Mexico [32], Malaysia [33], Pakistan [34], Philippines [35], Singapore [36], Taiwan [16], Sri Lanka [37], Costa Rica [38] and Vietnam [39]. Particle filtering is an estimation technique known as sequential Monte Carlo, which involves iteratively updating the estimated parameters using a sequence of observation. We used a number of new cases as observations, and the carrying capacity is the unknown parameter that is sequentially estimated.

The local mosquito populations are calculated using the life cycle equations, daily average temperature, and precipitation for each location around the globe with longitude and latitude resolutions of 0.25 degrees. The result is the time-dependent profile of the mosquito population. The calculated values are not the exact populations, rather they reflect the variation of the mosquito population in that location through time.

3.1.2. Dengue Transmission Reproductive Number

The value of dengue's basic reproductive number R_0

$$R_0 = \frac{V}{N^H} \frac{\alpha^2 \phi^{H\to V} \phi^{V\to H} \gamma^V}{\mu^V \sigma^H (\mu^V + \gamma^V)} \tag{1}$$

depends on the ratio of mosquito to human population V/N^H, mosquito biting rate α, human-to-vector and vector-to-human transmission probabilities per infectious bite $\phi^{H\to V}, \phi^{V\to H}$, extrinsic incubation period in the mosquito γ^V, mosquito mortality rate μ^V, and infectious period of human σ^H. These parameters are temperature-dependent and can be found in [40], except for V/N^H. The R_0 is derived assuming constant temperature and using Markovian equations. When $R_0 < 1$, the viral transmission will result in fewer new cases than past cases, and the infection will burn out if maintained for sufficient time. Conversely, interpretation that the infection is going to die out when $R_0 < 1$ is only relevant if this condition is satisfied for a long period. However, if the temperature is almost constant for a period longer than the relevant periods in the transmission model, the R_0 can be used to quantify the severity of virus transmission. In other words, when $R_0 > 1$, an infected individual is expected to generate more infection in the population. For this reason, we utilize the value of R_0 to gauge the severity of virus transmission. The mosquito-to-human population ratio V/N^H is time-dependent and is affected by the carrying capacity of the environment. To estimate the carrying capacity per human, we fit the transmission model to epidemic dengue curves from different locations around the world with a given human population. For each location, we obtain a constant component for the carrying capacity and a parameter that determines the variation of the carrying capacity with precipitation. We use the distribution of these parameters to find a statistical model for the carrying capacity per human. Although this model may not give the exact value for the carrying capacity, we use it as an approximate input in the calculation of R_0. Our calculation shows that R_0 is almost proportional to the carrying capacity; therefore, the calculated R_0 provides a good approximation for the time-dependent profile of the exact reproductive number. Higher R_0 values indicate greater dengue virus transmission in an area; therefore, users can quickly understand where control measures need to be initiated and evaluate if they are working by observing changes in R_0.

3.1.3. Dengue Transmission Risk

To estimate dengue transmission risk, we use the dengue transmission reproductive number and vector population calculated using the mechanistic model as predictive variables. For each location, we calculate the one-week moving average of the variables and

apply them in a trained neural network classifier to obtain a time-dependent transmission risk. Assuming the historical presence of the mosquitoes and the virus, the risk factor is a value between zero and one corresponding to the least and the most favorable environmental condition, respectively, for dengue transmission. To train the model, we used dengue epidemic data from Brazil, Mexico, Malaysia, Pakistan, Philippines, Singapore, Taiwan, Sri Lanka, Costa Rica and Vietnam. For each one of the epidemic curves, we manually tagged periods of high and low transmission. If the number of new cases is small and comparable to the baseline for the location, we tag that period as zero, otherwise as one. In addition, we calculate the corresponding predictive variable as described above for each location and period. Using the tag values as the target variable and the corresponding predictive variables, we train the neural network classifier for risk evaluation.

In Figure 2, we show the data used for training as green and yellow dots, indicating low (0) and high (1) transmission, respectively. The neural network classifier is a nonlinear model that assigns a risk value between zero and one to points on the surface defined by the predictive variables V/N^h and R_0. Within the depicted image, regions displaying risk values exceeding 0.5 are indicated by the color red, while those with values below 0.5 are represented by the color blue.

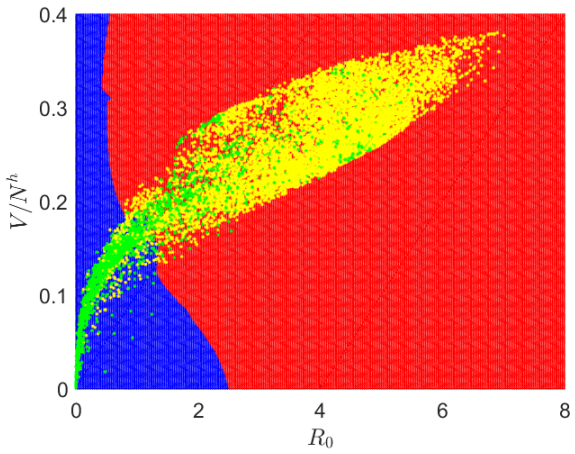


Figure 2. Dengue transmission risk map. Green and yellow dots represent training data; red and blue indicate the risk value above and below 0.5.

3.2. Forecasting Dengue Case Number

In this section, we outline the methodology we have developed for predicting the future incidence of ongoing dengue outbreaks. Our approach entails analyzing user-provided data, including reported incidence, location, and date, and extracting pertinent environmental information from our extensive database. These data are then utilized to generate forecasts, which are presented by the statistics including upper quantile, median, and lower quantile values of the forecast distribution. Case forecasts help PICTUREE–Aedes users understand how the dengue outbreak will change with time. Forecasts of increasing cases suggest that allocating addition resources to an area is needed due to continued transmission, whereas decreasing case forecasts suggest that resources can be reallocated and that the outbreak is getting smaller.

3.2.1. SEIR-SEI-EnKF Forecast

The SEIR-SEI-EnKF forecasts are generated by employing the disease transition model SEIR-SEI combined with the EnKF data-assimilation technique. The SEIR-SEI-EnKF model is a new approach for estimating and forecasting dengue outbreak dynamics, which has been detailed in [16]. The SEIR-SEI model is a compartmental model that describes the transmission of the dengue virus between the host and vector population groups. This model considers the compartments for the human population and vector population, such

as susceptible ($S^H and S^V$), exposed ($E^H and E^V$), infectious ($I^H and I^V$), and removed (R^H). Transitions between compartments are described by the following difference equations

$$X_{t+1} = f(X_t) = \begin{cases} S_t^H + \lambda^H P_t^H - \beta^{V\to H} S_t^H I_t^V - \mu^H S_t^H \\ E_t^H + \beta^{V\to H} S_t^H I_t^V - (\delta^H + \mu^H) E_t^H \\ I_t^H + \delta^H E_t^H - (\gamma^H + \mu^H) I_t^H \\ R_t^H + \gamma^H I_t^H - \mu^H R_t^H \\ S_t^V + \lambda^V P_t^V - \beta^{H\to V} S_t^V I_t^H - \mu^V S_t^V \\ E_t^V + \beta^{H\to V} S_t^V I_t^H - (\delta^V + \mu^V) E_t^V \\ I_t^V + \delta^V E_t^V - \mu^V I_t^V \end{cases} \tag{2}$$

where $X_t = [S_t^H, E_t^H, I_t^H, R_t^H, S_t^V, E_t^V, I_t^V]^\intercal$, $P, \lambda, \mu, \beta, \delta, and \gamma$ represent the population, birth rate, death rate, infection rate, incubation rate, and recovering rate, and upper note H for host species, and V for vector species.

The Ensemble Kalman Filter (EnKF) is a data assimilation technique that can be used to estimate both the states and parameters of a system simultaneously. The EnKF maintains an ensemble of the state variables X_t and the time-varying parameters $\Phi_t = [\beta^{V\to H} \beta^{H\to V} \lambda^V]^\intercal$, which can be put together in a matrix $A_t \in R^{10\times N}$ where N is the number of ensembles

$$A_t = \begin{bmatrix} X_t^1 & X_t^2 & \dots & X_t^N \\ \Phi_t^1 & \Phi_t^2 & \dots & \Phi_t^N \end{bmatrix}. \tag{3}$$

At each time step, the EnKF first uses the model simulations to predict the prior, which is the expected state of the system

$$A_{t+1}^- = F(A_t) = \begin{cases} f(X_t) \\ g(\Phi_t) \end{cases} + q_t \tag{4}$$

where $g(\Phi_t)$ is the difference equation for the parameters, and q_t is the model noise. Therefore, the ensemble covariance for the prior is given by

$$\sigma_t^{pri} = \frac{1}{N-1}(A_t^- - \bar{A_t^-})(A_t^- - \bar{A_t^-})^\intercal, \tag{5}$$

where $\bar{A_t^-}$ is the mean of ensemble members.

To update the estimated state and parameters of the system, the EnKF assimilates new observations y_t into the ensemble $\hat{Y}_t = y_t + \epsilon_t$, where ϵ_t is the perturbation vector. The assimilation step involves comparing the predicted states of the ensemble with the observed data and adjusting the estimated state and parameters of the system to better fit the observations

$$A_t = A_t^- + K_t(\hat{Y}_t - H_t A_t^-) \tag{6}$$

where $K_t = \sigma_t^{pri} H_t (H_t \sigma_t^{pri} H_t^\intercal + \sigma_t^{obs})^{-1}$ is the Kalman gain, and H_t is the measurement index matrix, $\sigma_t^{obs} = \frac{\epsilon_t \epsilon_t^\intercal}{N-1}$. Finally, the EnKF generates a new ensemble of model simulations by perturbing the updated estimated state and parameters of the system, and the process repeats for the next time step. In summary, the EnKF utilizes a data assimilation approach that combines real-time observations with mathematical models to accurately forecast dengue epidemics, helping to inform proactive measures for disease control and prevention.

3.2.2. Neural Network Forecast

The second approach to predict weekly cases uses a recurrent neural network model, as shown in Figure 3. Assuming that the index of the current week is represented by t, the model uses a sequence of weekly new cases up to t, $c_{t-4}, c_{t-3}, c_{t-2}, c_{t-1}, c_t$, as input to predict new cases in the next three weeks. In addition to the sequence of new cases,

the model uses the vector population and reproductive number sequences, calculated independently using the model discussed in the previous section.

The vector population $V_{t+h-4}, V_{t+h-3}, V_{t+h-2}, V_{t+h-1}, V_{t+h}$ and reproductive number $\{R_{t+h-4}, R_{t+h-3}, R_{t+h-2}, R_{t+h-1}, R_{t+h}\}$ include estimated vector population and reproductive number in the weeks ahead. Since the calculation of these sequences requires temperature as input, we use the average of temperature for the corresponding period in the past years. In addition to the sequences we discussed, we feed the error of the latest two predictions to the neural network model. We train this neural network model using historical dengue epidemic curves from different locations around the world. Although it is possible to train a neural network model that generates one, two, and three weeks of predictions, we choose to train independent models for each one of these prediction horizons. Here, the obtained output is a single value for each prediction time interval.

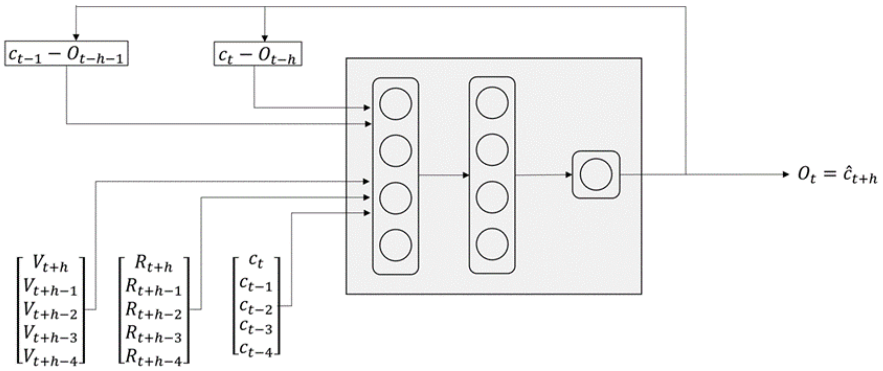


Figure 3. Neural network model for predicting dengue cases.

3.2.3. Particle Filter Forecast

As a third approach, we apply particle filter prediction to the transmission model, as depicted in Figure 4. In the transmission model, we divide the adult female mosquitoes into three compartments based on the virus development in the mosquitoes. We assume that all the newly developed adult mosquitoes are virus-free, (Adult-S). During the gonotrophic period, whose length is temperature-dependent, the mosquitoes may bite an infectious human, acquire the virus, and remain in the adult-exposed (Adult-E) state until they become infectious and are able to transmit the virus. The sojourn time for the Adult-E state is temperature-dependent. After becoming infectious, the mosquitoes infect susceptible humans during blood meals in their gonotrophic cycle until they die. We assume that the number of infecting bites per cycle is constant, and the bites occur uniformly in each cycle period. The human infection rate is $A_I B/(CN)$, where A_I is the population of infectious mosquitoes, B is the number of infecting bites per gonotrophic cycle, C is the cycle length, and N is the human population. Similarly, the mosquitoes' infection rate is $H_I B/(CN)$, where H_I is the infectious human population. In the particle filter estimation, we assume two unknown parameters, which are carrying capacity and the number of infecting bites per gonotrophic cycle. Like the mosquito life cycle model as shown in Figure 1, we account for non-Markovian transitions in the virus transmission model. Particle filtering is a sequential Monte Carlo approach that employs a group of samples to represent the posterior distribution of the unknown parameters. In each step of the process, a weight is assigned to each sample based on the observed data. During the resampling phase, particles with low weights are substituted by new particles in close proximity to the particles with higher weights. An advantage of the particle filter is its ability to handle observations from any probability distribution. In our particle filter estimation, we provide the algorithm with a sequence of weekly new human cases and estimate posterior distributions of new human cases for the one-week up to three-week horizon.

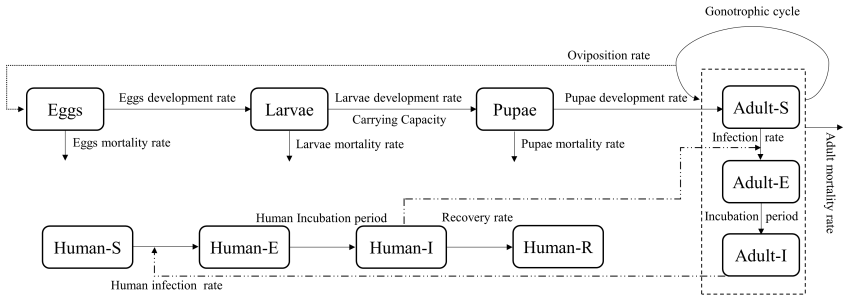


Figure 4. Particle filter transmission model for forecasting dengue cases.

3.2.4. Super Ensemble Forecast

Using the competing forecasts described above, a super-ensemble forecast can be created by applying a weighted average of the individual forecasts. The weights of the three individual forecasting methods is determined based on the root-mean-square error (RMSE) of those forecasts over prior predictions. Specifically, the RMSE of the forecasts made for time $t-3, t-2$, and $t-1$ can be used to generate the super-ensemble forecast for time T. The RMSE of each individual forecast $e^i (i = 1, 2, 3)$ can be calculated as follows:

$$e_t^i = \sqrt{\frac{1}{3}\sum_{\tau=t-3}^{t-1}(f_\tau^i - O_\tau)^2}, \tag{7}$$

where f_t^i is the median forecast of method i at time t, and O_t is the observation at time t. The weight of method i can be obtained by

$$w_t^i = \frac{1}{e_t^i \sum_{k=1}^{3} \frac{1}{e_t^k}}. \tag{8}$$

The super-ensemble forecast is given by

$$f_t^{se} = w_t^1 f_t^1 + w_t^2 f_t^2 + w_t^3 f_t^3. \tag{9}$$

4. System Architecture

The web application consists of several components as shown in Figure 5, which we broadly categorize into (1) the frontend client application and (2) the server backend.

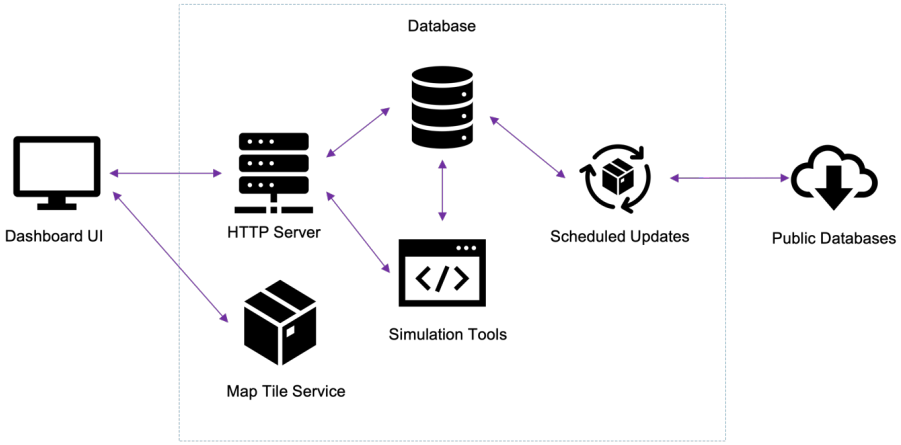


Figure 5. Components and architecture of the application.

4.1. Client Application

The web-based client is built on top of HTML and is accessible using any web browser. The application's graphical interface consists of an interactive map viewer and tools for data selection, filtering, parameter inputs, and submission of on-demand analyses. The map is displayed using the open-source JavaScript library leaflet.js [41]. The application communicates directly with the server backend to load maps, results, and data.

4.2. Server Backend

- HTTP server: The backend system exposes multiple API (application programming interface) endpoints to the client via an HTTP server. This HTTP server runs in the Node.js environment [42].
- Database: The database service consists of PostgreSQL and PostGIS [43] extensions to handle spatial queries. The database contains multiple tables to store data collected from various sources and simulation results.
- Map tile service: When the client requests the home page of the map tool, the HTTP server loads map tiles from a separate map tile service called MapTiler [44]. This service helps with memory management by loading only the parts and zoom levels required by the client. The service takes input from large tile files collected from OpenStreetMap [45].
- Simulation tools: The system can deploy and run executable code to perform on-demand analysis and simulations on selected data. The HTTP service directly communicates with the executables to pass arguments and to collect results.
- Scheduled updates: The system contains scripts that run regularly to update the database tables with new weather data. The script connects to the NOAA-NCEI [24] servers for daily weather observation data.

5. Results

5.1. Data Visualization

Figure 6 displays the graphical interface of the map tool, where the 2D projection map of the Earth is centrally located and where users are able to pan and zoom to their areas of interest. The interface's left side features various tools that assist with data selection and visualization. A comprehensive guide to the description and usage of each tool is available in the PICTUREE–Aedes Users Guide in the Supplementary Material.

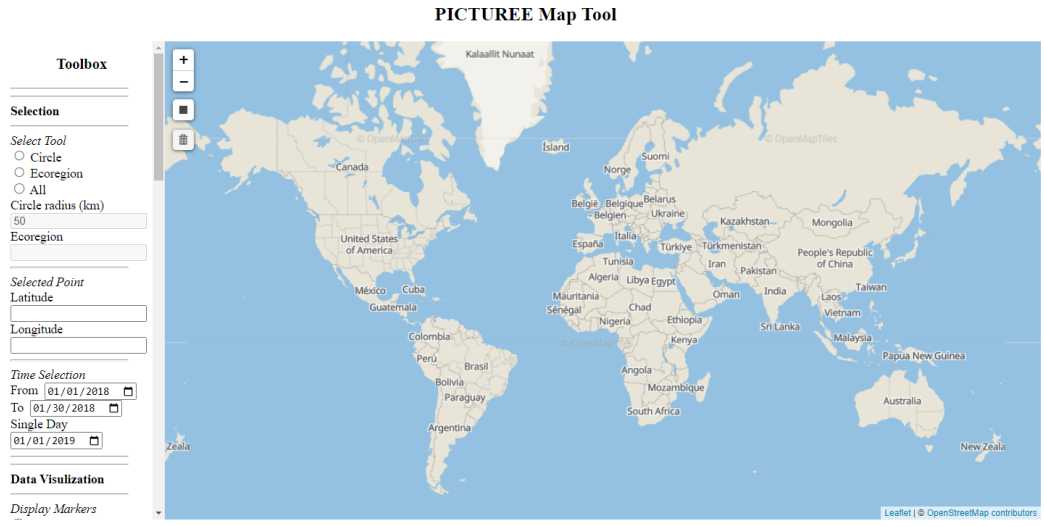
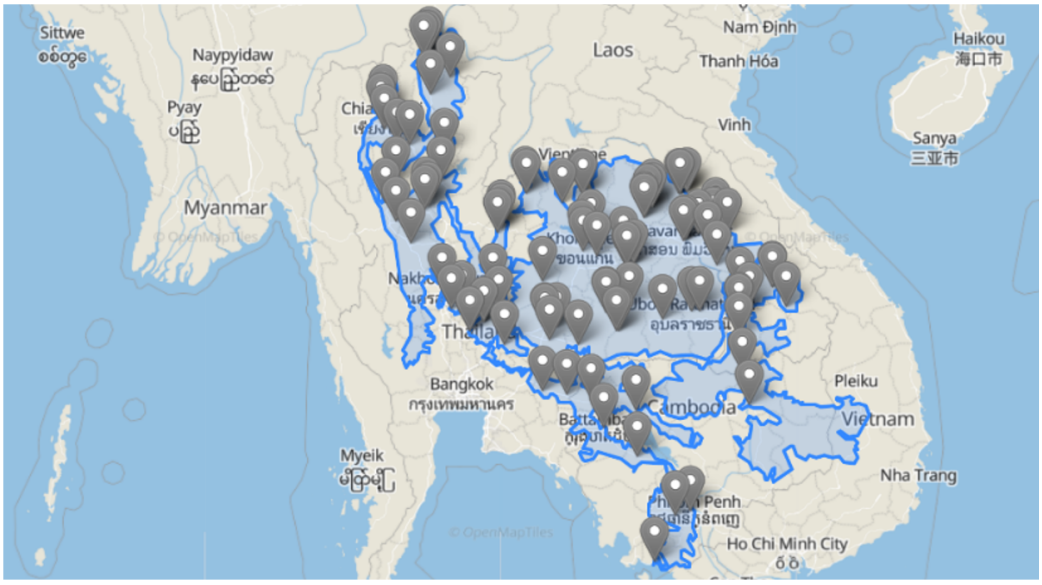


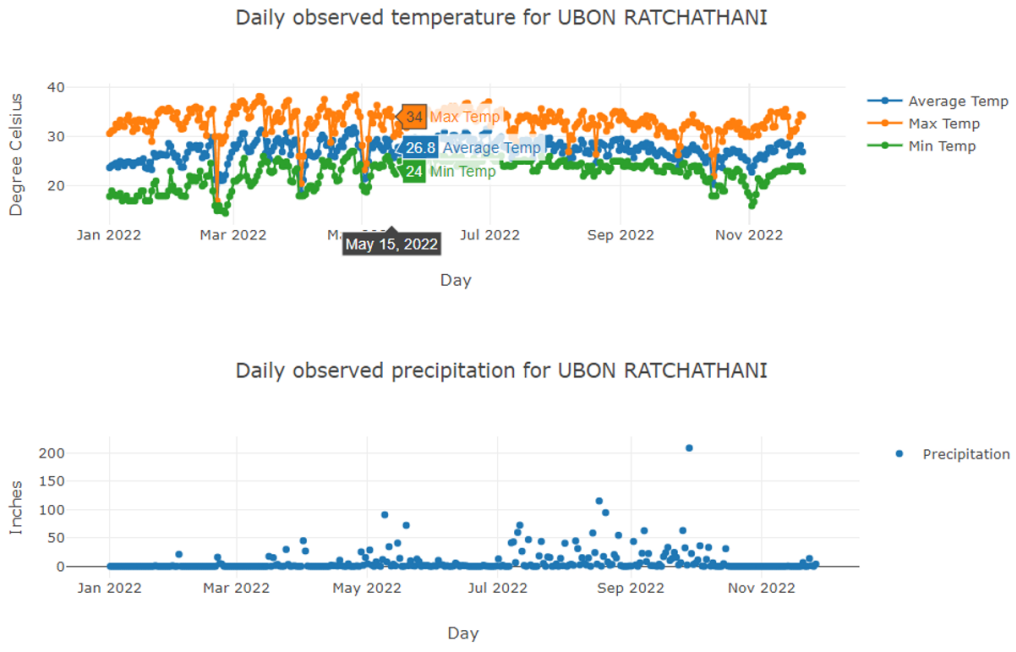
Figure 6. Graphical interface of the PICTUREE–Aedes.

The PICTUREE–Aedes enables users to identify weather stations and their recorded data based on their preferences. Figure 7a illustrates the distribution of weather stations in Central Indochina (Ecoregion). Users can select a station and plot its precipitation and

temperature data. Figure 7b demonstrates the daily temperature and precipitation data of station UBON RATCHATHANI throughout 2022.



(a)



(b)

Figure 7. Weather station data. (**a**) Locations of the weather stations in Central Indochina. (**b**) Temperature and precipitation data of the selected station in 2022.

Figure 8 presents the occurrence records of dengue, *Aedes aegypti*, and *Aedes albopictus*, marked, respectively, with red, blue, and purple dots in Central Indochina. Clicking on a dot will generate a popup that displays the country name and the year of occurrence.

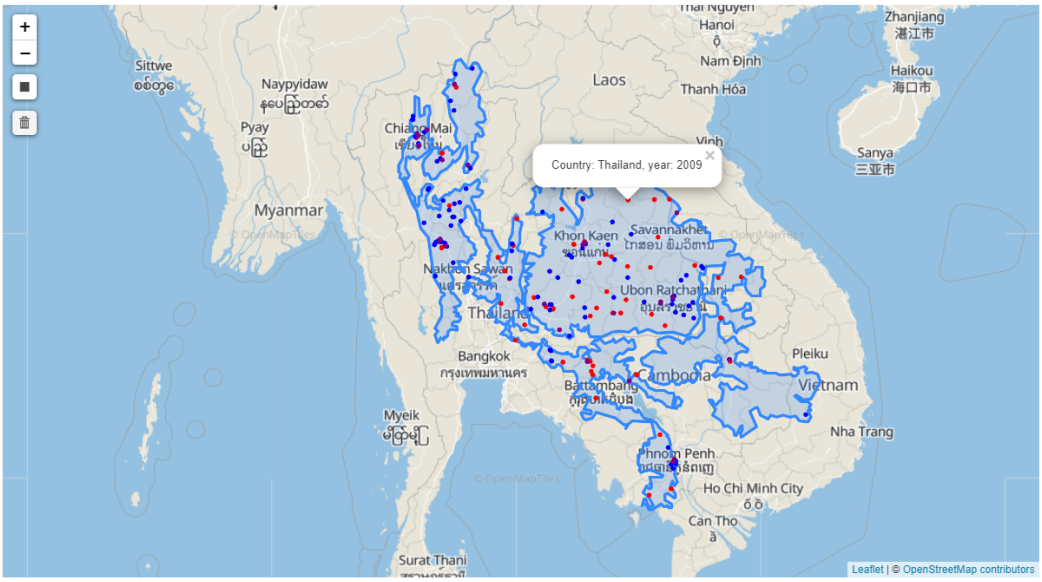


Figure 8. Occurrence data of dengue, *Aedes aegypti* mosquito, and *Aedes albopictus* mosquito.

Additionally, the PICTUREE–Aedes supports the display of information on Google Trends. Figure 9 illustrates the trend of "dengue" and "fever" in Cambodia during January 2018.

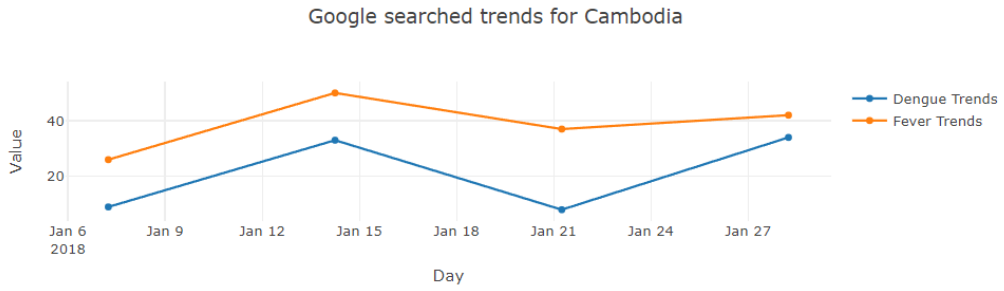


Figure 9. Google search trends of "dengue" and "fever".

5.2. Simulation Results

The PICTUREE–Aedes offers four types of mosquito simulation results, including daily average temperature, vector host ratio, reproduction number, and dengue risk, which can be visualized spatially and temporally. To view the data spatially, users must select a specific day, and a heat map of the selected area will display the selected data value. Figure 10 displays the average temperature (a), dengue transmission reproduction number (b), vector–host ratio (c), and dengue transmission risk (d) of the Indochina peninsula on 15 September 2022. Visualizing data beyond the next six months from the current time is limited.

To visualize data temporally, users must select a location and a time range. Figure 11 displays mosquito simulation data from 1 January to 11 June 2022 at the user-specified location. Please note that the query time frame is limited to a six-month horizon from the current date.

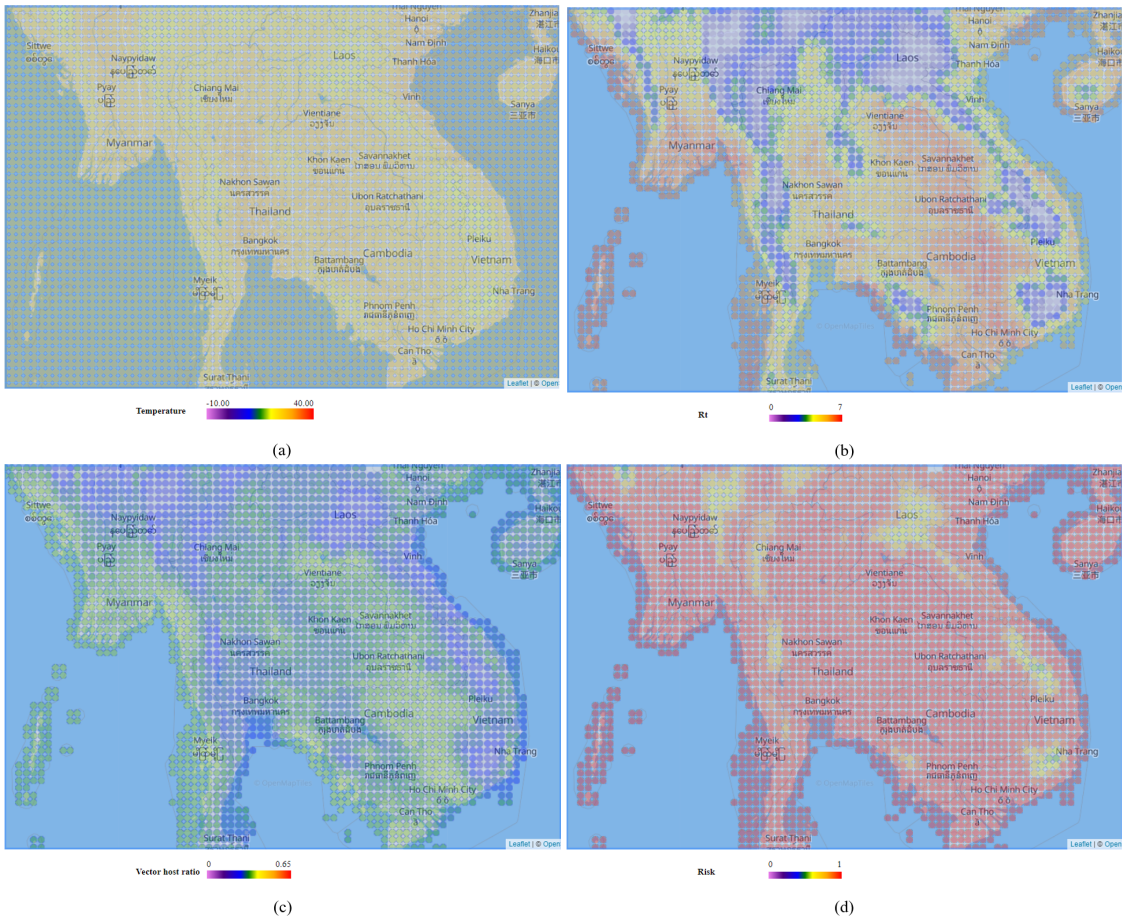


Figure 10. Mosquito simulation results visualized spatially. (**a**) Reanalyzed average temperature, range [−10, 40]. (**b**) Dengue transmission reproduction number R_t, range [0, 7]. (**c**) Vector–host ratio, range [0, 0.65]. (**d**) Dengue transmission risk, range [0, 1].

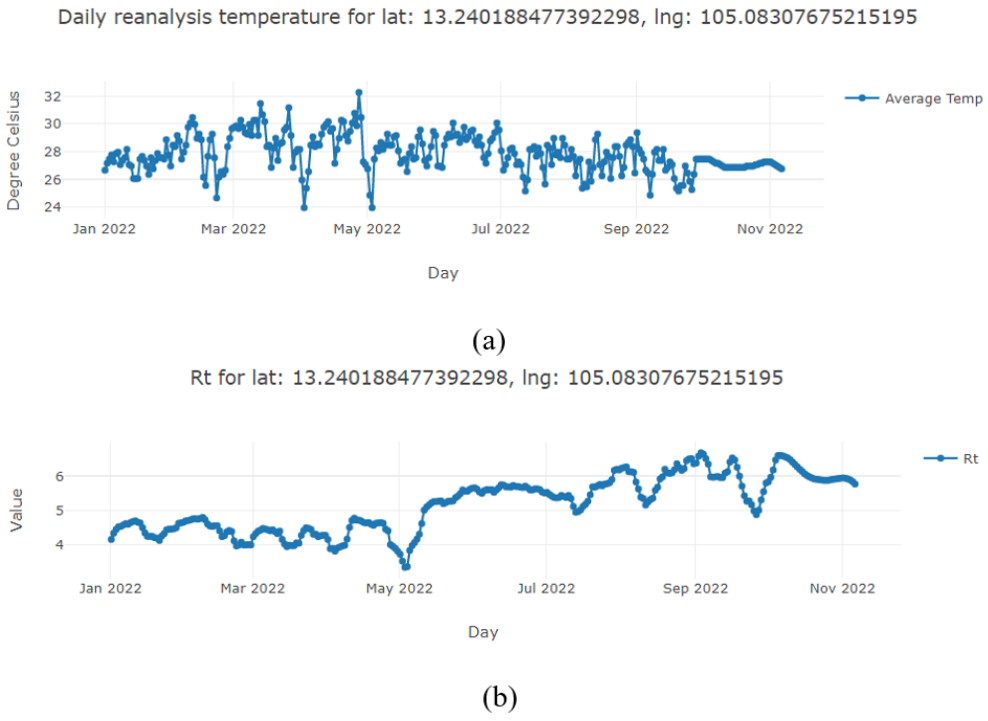
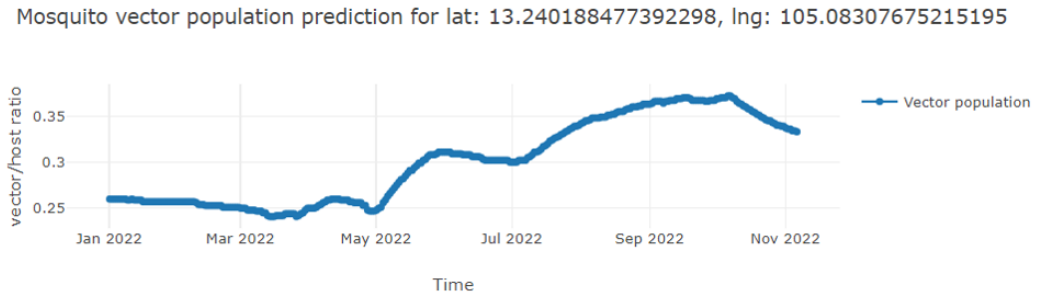
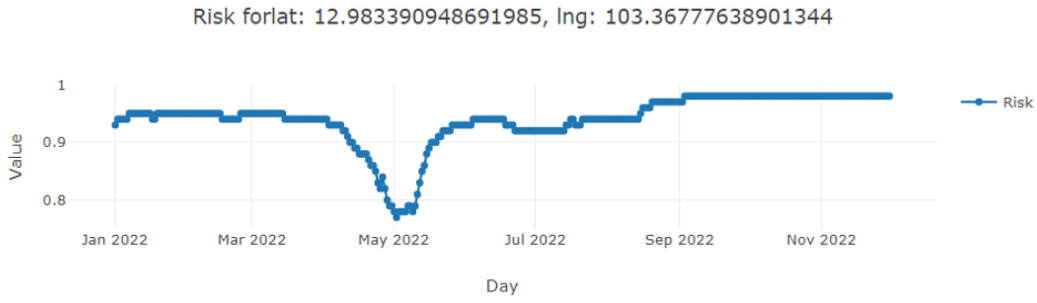


Figure 11. *Cont.*



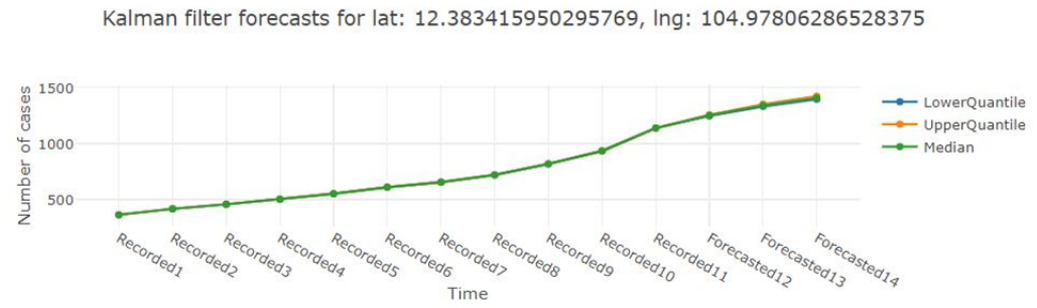
(c)



(d)

Figure 11. Mosquito simulation results visualized temporally. (**a**) Reanalyzed average temperature. (**b**) Dengue transmission reproduction number. (**c**) Vector–host ratio. (**d**) Dengue transmission risk.

The PICTUREE–Aedes provides three types of predictive tools and their ensemble forecast. Figure 12 illustrates the predicting results of the dengue outbreak in Cambodia during 2022. The first 13 data points (before 27 April) represent weekly dengue cases, while the remaining points (including/after 27 April) denote the predictions. Figure 12a displays the distribution of SEIR-SEI-EnKF predictions for the next three weeks, including quantiles and median. Figure 12b illustrates the neural network prediction for the next three weeks. Figure 12c shows the results of the particle filter, presenting the distribution for the next four weeks. Figure 12d displays the result of the super-ensemble, showcasing the median prediction from each member, the average, and weighted average of the members.



(a)

Figure 12. *Cont.*

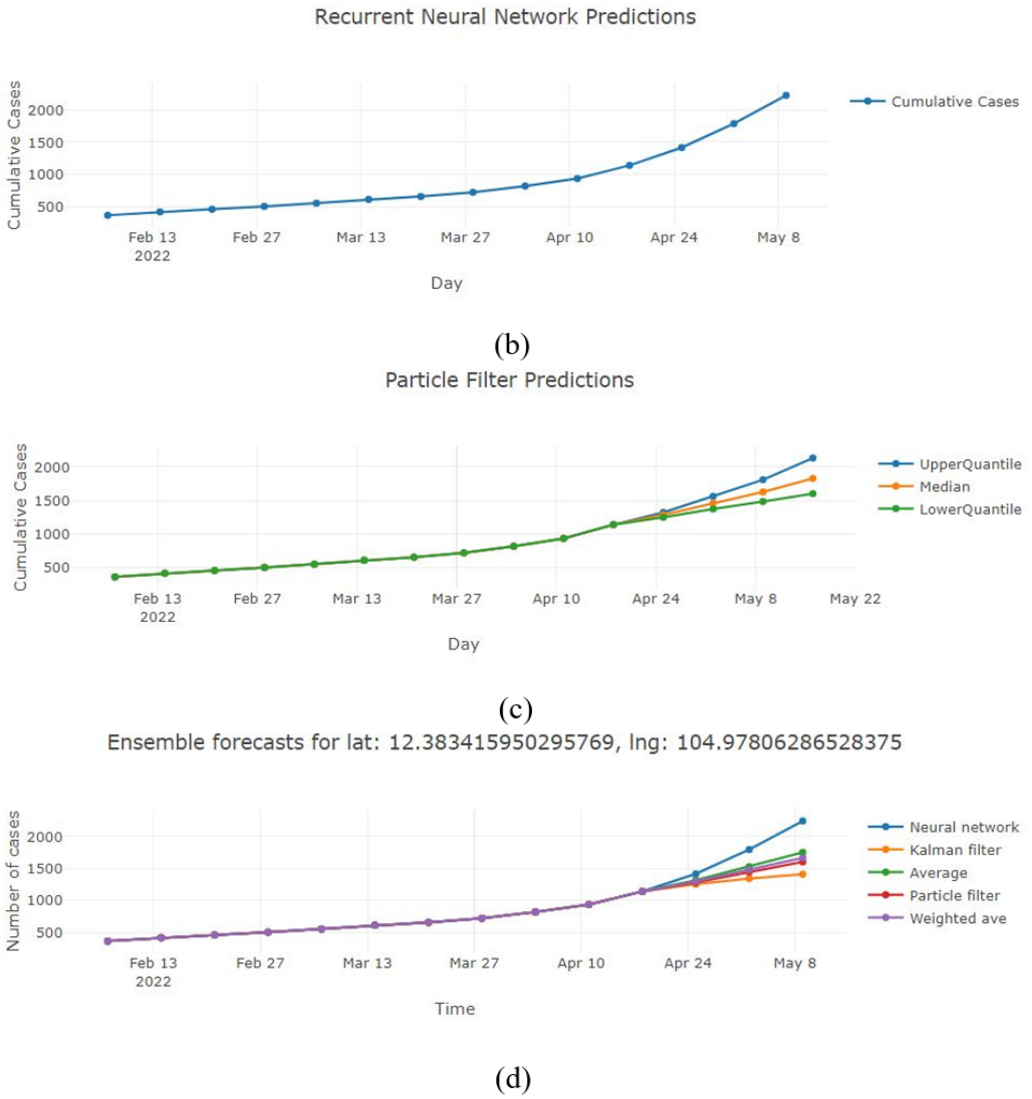


Figure 12. Forecasts of the dengue cases in Cambodia. (**a**) Forecasts of the SEIR-SEI-EnKF model. (**b**) Forecasts of the neural network model. (**c**) Forecasts of the particle filter model. (**d**) Forecasts of the super-ensemble model.

Table 2 displays the outcome of the prediction as well as the reported cases. All of the reported cases in the following weeks are situated within the prediction range, which confirms the precision of the prediction tools.

Table 2. Number of predicted cumulative cases and relative absolute error in Cambodia.

	27 April 2022	**4 May 2022**	**11 May 2022**
Reported cases	1379	1685	1995
Kalman filter (Median)	1229/10.9%	1293/23.3%	1338/32.9%
Neural network	1414/2.5%	1795/6.5%	2243/12.4%
Particle filter (Median)	1298/5.9%	1492/11.5%	1686/15.5%
Super ensemble (weighted average)	1302/5.6%	1500/11.0%	1703/14.6%

6. Discussion

Dengue fever is a viral mosquito-borne disease transmitted predominately by the *Aedes* mosquitoes. It infects an estimated 400 million people annually, with nearly half the world's population at risk of infection [46]. Dengue cases are influenced by complex interactions of ecology, environment, meteorological factors, and virus factors (serotype), among many others [47].

The PICTUREE-Aedes is an innovative and unique system that serves as a valuable resource for both researchers and health officials in their efforts to advance scientific research and to effectively manage dengue outbreaks. By collecting and aggregating a comprehensive set of critical factors related to dengue transmission, such as reanalyzed temperature, precipitation, occurrences of *Aedes* mosquitoes, and dengue cases, the PICTUREE-Aedes provides researchers with a rich dataset for their investigations. This data-driven approach allows researchers to gain deeper insights into the complex dynamics of dengue transmission and aids in the development of more effective prevention and control strategies. In addition to its research-oriented features, the PICTUREE-Aedes offers essential information to health officials, enabling them to proactively prepare and adapt to the changing situation during dengue outbreaks. The system employs sophisticated algorithms to estimate *Aedes* mosquito abundance, dengue transmission reproduction number, and dengue risk, providing crucial insights into the intensity and geographical spread of the disease. Armed with this knowledge, health officials can make informed decisions about when and where to implement targeted vector control measures, effectively minimizing the risk of dengue transmission and safeguarding public health. Moreover, the PICTUREE-Aedes goes beyond monitoring and estimation by offering valuable and accurate forecasting capabilities. By utilizing its advanced modeling techniques, the system can predict future dengue case counts, empowering health officials and healthcare facilities to deploy medical personnel and to allocate treatment resources more efficiently. This forecasting functionality is instrumental in anticipating the potential impact of dengue outbreaks, enabling proactive measures to be taken in terms of resource allocation, capacity planning, and public health awareness campaigns.

The selection of a 6-month forecast horizon was made without a specific rationale. The forecast horizon relies on the availability of dependable environmental variables for the future. Presently, we employ the average temperatures from previous years for the corresponding timeframe in the next six months. However, the current values of the calculated variables are adjusted based on the most recent available environmental data. The current levels will influence the future levels of the calculated variable. However, for the dengue case number forecast horizon, we chose 3 weeks because human intervention can change the underlying physical transmission process in the long term.

Concerning possible limitations of the proposed system, the temperatures displayed are collected by the weather stations; thus, they do not always reflect temperature in areas close to the weather station. For example, shaded and barren areas near the same weather station will be considered at the same daily temperature. In contrast, the actual temperature of those places can range significantly. There are many water sources beyond precipitation, and *Aedes* mosquitoes are very efficient at finding these areas to lay eggs, such as domestic or industrial wastewater, sprinkler systems, air conditioning units, etc. [48]. Therefore, the accuracy of the local temperature and precipitation and the estimation results obtained considering only temperature and precipitation can be erroneous in some areas, and local data are needed to adjust the models. Furthermore, human behaviors cannot be accounted for at this time [49]. Water collection during the dry seasons, litter accumulating water, and irrigating lawns have all been demonstrated to impact *Ae. aegytpi* numbers, and this global model cannot account for local behaviors at this time [50]. In the future, local models can be generated and placed within the global model. Additionally,there is an increasing significance of female mosquito vertical transmission through the dry quiescent egg population. These infected eggs play a potential role in triggering dengue incidence following periods of low mosquito activity, such as the dry season [51]. In our future work,

we plan to explore the dynamics of female mosquito vertical transmission in greater depth. This will involve investigating the factors influencing the rate of infected eggs, studying the survival and viability of infected offspring, and analyzing the potential amplification of dengue transmission through this mode. For the predictions, four different approaches provide reasonably accurate forecasts. However, some manual calibrations of the tools are needed to make the predictions more accurate at a local scale.

Future improvements include increasing the accuracy of the population abundance estimation in areas where no data are available, exploring the dynamics of female mosquito vertical transmission, optimizing the layout of the webpage, extrapolating the mosquito population dynamics from surveyed areas to unknown regions, and extending the tool to *Culex* mosquito genera.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pathogens12060771/s1.

Author Contributions: Conceptualization, C.Y., A.V., T.F., L.W.C. and C.S.; methodology, C.Y., A.V., T.F., L.W.C. and C.S.; software, T.F., C.Y. and A.V.; validation, C.Y., A.V., T.F., L.W.C. and C.S.; formal analysis, C.Y., A.V., T.F., L.W.C. and C.S.; investigation, C.Y., A.V., T.F., L.W.C. and C.S.; resources, C.Y., A.V., T.F., L.W.C. and C.S.; data curation, A.V. and T.F.; writing—original draft preparation, C.Y., A.V. and T.F.; writing—review and editing, C.Y., A.V., T.F., L.W.C. and C.S.; visualization, T.F.; supervision, L.W.C. and C.S.; project administration, L.W.C. and C.S.; funding acquisition, C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Department of the Army, US Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Fort Detrick, under DWFP grant W911QY-19-1-0004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available. Weather station data can be found here: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516 (accessed on 20 April 2023). ERA5 hourly data are openly available in "ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)", DOI: 10.24381, reference number [25]. *Aedes* mosquito occurrence data are openly available in "Kraemer M. U. G. 2015. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.47v3c", reference number [29]. Dengue occurrence data is openly available in "Messina, J., Brady, O., Pigott, D. et al. A global compendium of human dengue virus occurrence. Sci Data 1, 140004 (2014). https://doi.org/10.1038/sdata.2014.4", reference number [30]. Google trends data can be found here: https://trends.google.com/trends/explore?date=now%201-d&geo=US&q=dengue,fever&hl=en (accessed on 20 April 2023). Other data are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wilder-Smith, A.; Ooi, E.E.; Horstick, O.; Wills, B. Dengue. *Lancet* **2019**, *393*, 350–363. [CrossRef] [PubMed]
2. Rocklöv, J.; Quam, M.B.; Sudre, B.; German, M.; Kraemer, M.U.; Brady, O.; Bogoch, I.I.; Liu-Helmersson, J.; Wilder-Smith, A.; Semenza, J.C.; et al. Assessing seasonal risks for the introduction and mosquito-borne spread of Zika virus in Europe. *EBioMedicine* **2016**, *9*, 250–256. [CrossRef] [PubMed]
3. Polwiang, S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003–2017). *BMC Infect. Dis.* **2020**, *20*, 208. [CrossRef]
4. Wu, P.C.; Guo, H.R.; Lung, S.C.; Lin, C.Y.; Su, H.J. Weather as an effective predictor for occurrence of dengue fever in Taiwan. *Acta Trop.* **2007**, *103*, 50–57. [CrossRef] [PubMed]
5. Yang, H.; Macoris, M.D.L.D.G.; Galvani, K.C.; Andrighetti, M.T.M.; Wanderley, D.M.V. Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. *Epidemiol. Infect.* **2009**, *137*, 1188–1202. [CrossRef]
6. Watts, D.M.; Burke, D.S.; Harrison, B.A.; Whitmire, R.E.; Nisalak, A. *Effect of Temperature on the Vector Efficiency of Aedes aegypti for Dengue 2 Virus*; Army Medical Research Inst of Infectious Diseases: Fort Detrick, MD, USA, 1986.
7. Struchiner, C.J.; Rockloev, J.; Wilder-Smith, A.; Massad, E. Increasing dengue incidence in Singapore over the past 40 years: Population growth, climate and mobility. *PLoS ONE* **2015**, *10*, e0136286. [CrossRef]

8. Focks, D.A.; Haile, D.G.; Daniels, E.; Mount, G.A. Dynamic life table model for Aedes aegypti (Diptera: Culicidae): Analysis of the literature and model development. *J. Med. Entomol.* **1993**, *30*, 1003–1017. [CrossRef]
9. Riad, M.H.; Cohnstaedt, L.W.; Scoglio, C.M. Risk Assessment of Dengue Transmission in Bangladesh Using a Spatiotemporal Network Model and Climate Data. *Am. J. Trop. Med. Hyg.* **2021**, *104*, 1444. [CrossRef]
10. Eisen, L.; Monaghan, A.J.; Lozano-Fuentes, S.; Steinhoff, D.F.; Hayden, M.H.; Bieringer, P.E. The impact of temperature on the bionomics of Aedes (Stegomyia) aegypti, with special reference to the cool geographic range margins. *J. Med. Entomol.* **2014**, *51*, 496–516. [CrossRef]
11. Couret, J.; Dotson, E.; Benedict, M.Q. Benedict. Temperature, larval diet, and density effects on development rate and survival of Aedes aegypti (Diptera: Culicidae). *PLoS ONE* **2014**, *9*, e87468. [CrossRef]
12. Barbazan, P.; Guiserix, M.; Boonyuan, W.; Tuntaprasart, W.; Pontier, D.; Gonzalez, J.P. Modelling the effect of temperature on transmission of dengue. *Med. Vet. Entomol.* **2010**, *24*, 66–73. [CrossRef] [PubMed]
13. Martinez, E.Z.; Silva, E.A.S.D.; Fabbro, A.L.D. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Rev. Soc. Bras. Med. Trop.* **2011**, *44*, 436–440. [CrossRef] [PubMed]
14. Brooks, L.C.; Farrow, D.C.; Hyun, S.; Tibshirani, R.J.; Rosenfeld, R. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS Comput. Biol.* **2015**, *11*, e1004382. [CrossRef]
15. Chen, Y.; Liu, T.; Yu, X.; Zeng, Q.; Cai, Z.; Wu, H.; Zhang, Q.; Xiao, J.; Ma, W.; Pei, S.; et al. An ensemble forecast system for tracking dynamics of dengue outbreaks and its validation in China. *PLoS Comput. Biol.* **2022**, *18*, e1010218. [CrossRef]
16. Yi, C.; Cohnstaedt, L.W.; Scoglio, C.M. SEIR-SEI-EnKF: A new model for estimating and forecasting dengue outbreak dynamics. *IEEE Access* **2021**, *9*, 156758–156767. [CrossRef]
17. DeFelice, N.B.; Little, E.; Campbell, S.R.; Shaman, J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. *Nat. Commun.* **2017**, *8*, 14592. [CrossRef]
18. ur, Rehman, M.J.; Dass, S.C.; Asirvadam, V.S.; Adly, A. Parameter estimation for nonlinear disease dynamical system using particle filter. In Proceedings of the 2015 International Conference on Computer, Control, Informatics and Its Applications (IC3INA), Bandung, Indonesia, 5–7 October 2015; pp. 143–147.
19. Zhao, X.; Li, K.; Ang, C.K.E.; Cheong, K.H. A deep learning based hybrid architecture for weekly dengue incidences forecasting. *Chaos Solitons Fractals* **2023**, *168*, 113170. [CrossRef]
20. Ferdousi, T.; Cohnstaedt, L.W.; Scoglio, C.M. A Windowed Correlation-Based Feature Selection Method to Improve Time Series Prediction of Dengue Fever Cases. *IEEE Access* **2021**, *9*, 141210–141222. [CrossRef]
21. Xu, J.; Xu, K.; Li, Z.; Meng, F.; Tu, T.; Xu, L.; Liu, Q. Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *Int. J. Environ. Res. Public Health* **2020**, *17*, 453. [CrossRef]
22. Deb, S.; Deb, S. An ensemble method for early prediction of dengue outbreak. *J. R. Stat. Soc. Ser. Stat. Soc.* **2022**, *185*, 84–101. [CrossRef]
23. Yamana, T.K.; Kandula, S.; Shaman, J. Superensemble forecasts of dengue outbreaks. *J. R. Soc. Interface* **2016**, *13*, 20160410. [CrossRef]
24. NOAA National Centers for Environmental Information (2001): Global Surface Summary of the Day—GSOD. NOAA National Centers for Environmental Information. Available online: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516 (accessed on 3 April 2023).
25. Hersbach, H.; Bell, B.; Berrisford, P.; Biavati, G.; Horányi, A.; Muñoz Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Rozum, I.; et al. ERA5 hourly data on single levels from 1979 to present. *Copernic. Clim. Chang. Serv. (C3S) Clim. Data Store (CDS)* **2018**, *10*, 10.24381.
26. Olson, D.M.; Dinerstein, E. The Global 200: Priority ecoregions for global conservation. *Ann. Mo. Bot. Gard.* **2002**, *89*, 199–224. [CrossRef]
27. The Nature Conservancy; USDA Forest Service; U.S. Geological Survey. Based on Bailey, Robert G. In *Description of the Ecoregions of the United States*, 2nd ed.; Misc. Pub. No. 1391, Map Scale 1:7,500,000; USDA Forest Service: Washington, DC, USA, 1995; 108p.
28. Wiken, E.B. Terrestrial ecozones of Canada. In *Ecological Land Classification Series No. 19*; Environment Canada, Hull Quebec: Gatineau, QC, Canada, 1986; p. 26.
29. Kraemer, M.U.; Sinka, M.E.; Duda, K.A.; Mylne, A.; Shearer, F.M.; Brady, O.J.; Messina, J.P.; Barker, C.M.; Moore, C.G.; Carvalho, R.G.; et al. The global compendium of Aedes aegypti and Ae. albopictus occurrence. *Sci. Data* **2015**, *2*, 1–8.
30. Messina, J.P.; Brady, O.J.; Pigott, D.M.; Brownstein, J.S.; Hoen, A.G.; Hay, S.I. A global compendium of human dengue virus occurrence. *Sci. Data* **2014**, *1*, 1–6. [CrossRef] [PubMed]
31. Dengue Fever-Explore-Google Trends. Available online: https://trends.google.com/trends/explore?date=now%201-d&geo=US&q=dengue,fever&hl=en (accessed on 3 April 2023).
32. PAHO/WHO Data—National Dengue Fever Cases. Available online: https://www3.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html (accessed on 15 May 2023).
33. Murphy, A.; Rajahram, G.S.; Jilip, J.; Maluda, M.; William, T.; Hu, W.; Reid, S.; Devine, G.J.; Frentiu, F.D. Incidence and epidemiological features of dengue in Sabah, Malaysia. *PLoS Negl. Trop. Dis.* **2020**, *14*, e0007504. [CrossRef]
34. Chaudhry, M.; Ahmad, S.; Rashid, H.B.; Uddin, I. Dengue epidemic in postconflict swat district, Khyber Pakhtunkhwa, Pakistan, 2013. *Am. J. Trop. Med. Hyg.* **2017**, *96*, 899. [CrossRef]

35. Xu, Z.; Bambrick, H.; Yakob, L.; Devine, G.; Frentiu, F.D.; Salazar, F.V.; Bonsato, R.; Hu, W. High relative humidity might trigger the occurrence of the second seasonal peak of dengue in the Philippines. *Sci. Total Environ.* **2020**, *708*, 134849. [CrossRef]
36. Hii, Y.L.; Zhu, H.; Ng, N.; Ng, L.C.; Rocklöv, J. Forecast of dengue incidence using temperature and rainfall. *PLoS Negl. Trop. Dis.* **2012**, *6*, e1908. [CrossRef]
37. Liyanage, P.; Tissera, H.; Sewe, M.; Quam, M.; Amarasinghe, A.; Palihawadana, P.; Wilder-Smith, A.; Louis, V.R.; Tozan, Y.; Rocklöv, J. A spatial hierarchical analysis of the temporal influences of the El Nino-southern oscillation and weather on dengue in Kalutara District, Sri Lanka. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1087. [CrossRef]
38. Vásquez, P.; Loría, A.; Sanchez, F.; Barboza, L.A. Climate-driven statistical models as effective predictors of local dengue incidence in costa rica: A generalized additive model and random forest approach. *Rev. MatemáTica TeoríA Apl.* **2020**, *27*, 1–21. [CrossRef]
39. Cuong, H.Q.; Hien, N.T.; Duong, T.N.; Phong, T.V.; Cam, N.N.; Farrar, J.; Nam, V.S.; Thai, K.T.; Horby, P. Quantifying the emergence of dengue in Hanoi, Vietnam: 1998–2009. *PLoS Negl. Trop. Dis.* **2011**, *5*, e1322. [CrossRef] [PubMed]
40. Lourenço, J.; Recker, M. The 2012 Madeira dengue outbreak: Epidemiological determinants and future epidemic potential. *PLoS Negl. Trop. Dis.* **2014**, *8*, e3083. [CrossRef] [PubMed]
41. Edler, D.; Vetter, M. The simplicity of modern audiovisual web cartography: An example with the open-source javascript library leaflet. js. *KN-J. Cartogr. Geogr. Inf.* **2019**, *69*, 51–62. [CrossRef]
42. Tilkov, S.; Vinoski, S. Node. js: Using JavaScript to build high-performance network programs. *IEEE Internet Comput.* **2010**, *14*, 80–83. [CrossRef]
43. Zimányi, E.; Sakr, M.; Lesuisse, A. MobilityDB: A mobility database based on PostgreSQL and PostGIS. *ACM Trans. Database Syst. (TODS)* **2020**, *45*, 1–42. [CrossRef]
44. Cascón-Katchadourian, J.D.; Ruiz-Rodríguez, A.Á. Description and Evaluation of MapTiler Software: From Scanned Map to Interactive Layer for Web Publishing. 2016. Available online: https://dialnet.unirioja.es/servlet/articulo?codigo=5730199 (accessed on 20 April 2023).
45. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
46. Züst, R.; Toh, Y.X.; Valdés, I.; Cerny, D.; Heinrich, J.; Hermida, L.; Marcos, E.; Guillén, G.; Kalinke, U.; Shi, P.Y.; et al. Type I interferon signals in macrophages and dendritic cells control dengue virus infection: Implications for a new mouse model to test dengue vaccines. *J. Virol.* **2014**, *88*, 7276–7285. [CrossRef]
47. Xiang, J.; Hansen, A.; Liu, Q.; Liu, X.; Tong, M.X.; Sun, Y.; Cameron, S.; Hanson-Easey, S.; Han, G.S.; Williams, C.; et al. Association between dengue fever incidence and meteorological factors in Guangzhou, China, 2005–2014. *Environ. Res.* **2017**, *153*, 17–26. [CrossRef]
48. World Health Organization. *Manual on Environmental Management for Mosquito Control, with Special Emphasis on Malaria Vectors*; World Health Organization: Geneva, Switzerland, 1982.
49. Barrera, R.; Amador, M.; MacKay, A.J. Population dynamics of Aedes aegypti and dengue as influenced by weather and human behavior in San Juan, Puerto Rico. *PLoS Negl. Trop. Dis.* **2011**, *5*, e1378. [CrossRef]
50. Nguyen, L.A.P.; Clements, A.C.; Jeffery, J.A.; Yen, N.T.; Nam, V.S.; Vaughan, G.; Shinkfield, R.; Kutcher, S.C.; Gatton, M.L.; Kay, B.H.; et al. Abundance and prevalence of Aedes aegypti immatures and relationships with household water storage in rural areas in southern Viet Nam. *Int. Health* **2011**, *3*, 115–125. [CrossRef] [PubMed]
51. Kurnia, N.; Kaitana, Y.; Salaki, C.L.; M.; ey, L.C.; Tuda, J.S.B.; Tallei, T.E. Study of Dengue Virus Transovarial Transmission in *Aedes* spp. in Ternate City Using Streptavidin-Biotin-Peroxidase Complex Immunohistochemistry. *Infect. Dis. Rep.* **2022**, *14*, 765–771. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.