SEIR-SEI-EnKF: a new model for estimating and forecasting dengue outbreak dynamics

CHUNLIN YI1, LEE W. COHNSTAEDT2, AND CATERINA M. SCOGLIO.1, (Senior Member, IEEE)

1Department of Electrical and Computer Engineering, College of Engineering, Kansas State University, Manhattan, KS 66506 USA
2Foreign Arthropod-Borne Animal Diseases Research Unit, National Bio and Agrodefense Facility, USDA ARS, Manhattan, KS, USA

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

This research has been supported by the Department of the Army, U.S. Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division, Fort Detrick (DWFP Grant W911QY-19-1-0004). 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.

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 estimated 360 million dengue infections per year [3]. In Singapore, large dengue outbreaks occurred almost every 5-6 years since the late 1980s despite 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 deterministic 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 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 cases 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 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 develop a real-time forecast system of hand-foot-and-mouth disease outbreaks in [12]. In addition, an 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 [22]. Those methods were the basic particle filter (PF), maximum likelihood estimation viaiterated filtering (MIF), 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 a twin experiment 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 an estimation 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 from epidemic week 27/2015 to 5/2016 [31], Singapore from epidemic week 1/2020 to 50/2020 [33] and Rio de Janeiro Brazil 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 was 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

1) **SEIR-SEI 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 (\( S_H \)); 2) exposed (\( E_H \)), defined as the population that are unexposed to the virus but can be infected by the bites from the infectious mosquitoes; 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, the transmission probability. The infection rates per susceptible human and susceptible mosquito are given by

\[
p_H \frac{N_M}{N_H} \frac{b}{\bar{N}_M} \frac{I_M}{N_H} = p_H \frac{b}{\bar{N}_H} \frac{I_M}{N_H} = \beta_H I_M
\]

\[p_M \frac{b}{\bar{N}_H} \frac{I_H}{N_H} = p_M \frac{b}{\bar{N}_M} \frac{I_H}{N_H} = \beta_M I_H
\]

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. \( \bar{I}_M = \frac{I_M}{N_M}, \bar{I}_H = \frac{I_H}{N_H} \) are the normalized values of the infectious population.

Transitions between compartments are described by (3).

\[
\begin{align*}
\bar{S}_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_M - \mu_H \bar{S}_H \\
\bar{E}_H(t) &= \beta_H \bar{S}_H \bar{I}_M - (\delta_H + \mu_H)(\bar{E}_H(t)) \\
\bar{I}_H(t) &= \delta_H \bar{E}_H - (\gamma_H + \mu_H) \bar{I}_H \\
\bar{R}_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H
\end{align*}
\]

Where \( \lambda_H \) denotes the birth rate for human; \( \mu \) denotes the death rate for both species; \( \delta \) denotes the transition rate from exposed to infectious; \( \gamma \) denotes the transition rate from infectious to removed; \( \bar{A} \) denotes the normalized mosquito recruitment 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 \( \delta_H \) and \( \gamma_H \) satisfy the following distributions \( \frac{1}{\bar{\delta}_H} \sim N(7,1.5), \frac{1}{\bar{\gamma}_H} \sim N(4.5,1.25) \); 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 \( \frac{1}{\bar{\mu}_M} \sim N(14,\sigma), \frac{1}{\bar{\gamma}_M} \sim N(10,\sigma). \)

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

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).

\[
\begin{align*}
\bar{S}_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_H - \mu_H \bar{S}_H \\
\bar{I}_H(t) &= \beta_H \bar{S}_H \bar{I}_H - (\gamma_H + \mu_H) \bar{I}_H \\
\bar{R}_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H
\end{align*}
\]

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

\[
\begin{align*}
\bar{S}_H(t) &= \lambda_H - \beta_H \bar{S}_H \bar{I}_M - \mu_H \bar{S}_H \\
\bar{E}_H(t) &= \beta_H \bar{S}_H \bar{I}_M - (\delta_H + \mu_H)(\bar{E}_H(t)) \\
\bar{I}_H(t) &= \delta_H \bar{E}_H - (\gamma_H + \mu_H) \bar{I}_H \\
\bar{R}_H(t) &= \gamma_H \bar{I}_H - \mu_H \bar{R}_H
\end{align*}
\]

The variables in (4) and (5) denote the same states or parameters as 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 use an efficient data assimilation algorithm – the Ensemble Kalman Filter (EnKF). The filtering technique of EnKF is based on Monte Carlo. The EnKF assumes a Gaussian distribution of both the prior and likelihood and adjusts the prior distribution to a posterior using Bayes rule.
The EnKF maintains an ensemble of the unknown parameters and the state variables. Let the state variables defined as \( x_k \in \mathbb{R} \), the unknown parameters \( \Phi_k \in \mathbb{R} \). In the ensemble presentation, they can be put together in a matrix \( \hat{A}_k \in \mathbb{R}^{N \times N} \), holding \( N \) ensemble member at time \( k \).

\[
\hat{A}_k = \begin{bmatrix}
    x_1^1 & x_1^2 & \ldots & x_1^N \\
    \phi_1^1 & \phi_2^1 & \ldots & \phi_N^1 \\
    \vdots & \vdots & \ddots & \vdots \\
    \phi_1^N & \phi_2^N & \ldots & \phi_N^N
\end{bmatrix}
\]  

(6)

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

\[
\hat{A}_{k+1} = f(\hat{A}_k) + q_k,
\]

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 \( \hat{A}_k \) is the best estimate of the true state. Therefore, the ensemble covariance for prior at time \( k \) \( \bar{P}_k \) is given by equation (8).

\[
\bar{P}_k = \frac{1}{N-1}(\bar{A}_k - \bar{\hat{A}}_k)(\bar{A}_k - \bar{\hat{A}}_k)^T
\]

(8)

The ensemble presentation of measurements can be formulated as

\[
Y_k = \begin{bmatrix}
    y_1^1 \\
    y_2^1 \\
    \vdots \\
    y_N^1
\end{bmatrix} = y_k + \epsilon_k,
\]

where \( y_k \) is the true measurement and \( \epsilon_k \) is the perturbation vector.

\[
y_k = H_k \hat{A}_k
\]

(10)

\( H_k \) is the measurement index matrix at time \( k \).

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

\[
V_k = \frac{\epsilon_k \epsilon_k^T}{N-1}.
\]

(11)

The Kalman gain is given by

\[
G_k = P_k^{-1} H_k (H_k P_k^{-1} H_k^T + V_k)^{-1}.
\]

(12)

The posterior and its covariance are then updated by

\[
\hat{A}_k = \bar{\hat{A}}_k + G_k (Y_k - H_k \bar{\hat{A}}_k),
\]

(13)

\[
P_k = \frac{1}{N-1}(\bar{A}_k - \bar{\hat{A}}_k)(\bar{A}_k - \bar{\hat{A}}_k)^T.
\]

(14)

In this analysis, the mean value of priors \( \hat{A}_k \) is our forecast at time \( k \) and the updated posterior is used to produce the forecast of states and the estimation of parameters in the next time step.

**D. TWIN EXPERIMENT**

The twin experiment is an effective method to verify the feasibility and evaluate the accuracy of parameters estimation techniques [32]. In the twin experiment, 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, a synthetic time series of dengue incidence and mosquito abundance is generated by an SEIR-SEI model with the parameters and states initial condition listed in table 1. In this simulation, the prescribed values for \( \beta_H, \beta_M \), and \( \gamma \) are piece-wise constant, which means their values have a step change during week 25 to week 26. This action is aimed to mimic the dynamic characteristics of these parameters. We add white noise (zero mean and 0.0001 variance) to the synthetic data and input it to EnKF as the observation of state \( R_H \) (dengue incidence) and the sum of \( S_M, E_M \) and \( I_M \) (mosquito abundance). The estimated parameters and unknown states by EnKF will be compared with the synthetic model to evaluate the performance of the EnKF.

**E. ESTIMATIONS AND FORECASTS OF SEIR-SEI-ENKF MODEL**

The weekly series of incidence data (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 simulates 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 assess the prediction accuracy. At the beginning of each simulation, each model states are initialized with a random value. Then the system starts to assimilate the observations. The simulation is replicated 200 times, the average estimations and the distribution of the predictions are presented in the results.

**F. ESTIMATIONS AND FORECASTS OF SEIR-SEI-ENKF MODEL**

Three disease models – SEIR-SIE, SEIR, and SRI model are applied to the EnKF for predicting the dengue cases in Kaohsiung Taiwan. The predictions and the assimilations 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 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 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 $\beta_H$ 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 $\beta_M$ has a large overshoot and oscillation 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 $A$ 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 $R_{IH}$ 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 $E_H$ and $I_H$. In addition, the estimated epidemic peak week (the week when $I_H$ reaches the maximum) matches the real one. The errors for peak value of states $E_H$ and $I_H$ are 20% and 13.5%. For mosquito population, the total number of female mosquitoes is observed. As seen in Fig.1(c), the estimations and predictions on states $S_M$ and $E_M$ are almost accurate. For state $I_M$, 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 – week 22), the infection rates for both species keep increasing to their peak value in week 21 (1.52 for $\beta_H$ and 5.65 for $\beta_M$); the mosquito abundance rate $A$ has two “steady states” – around $3.3 \times 10^{-4}$ before peak weeks and around $1.8 \times 10^{-4}$ after. The value of infection rate $\beta_M$ is much greater than that of $\beta_H$ that 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 100 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 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
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   |

human population are the week 21. The estimated normalized magnitude of the infectious population is $9.8 \times 10^{-4}$ 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 that corresponds to the high-incidence period for human population.

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

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 4 step prediction and the 5 step 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 confirmed cases are presented in Fig.4. The weekly confirmed cases is given by $N_{t+1}^H = R_{t+1}^H - R_{t}^H$, 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 adjust the model parameters and calculate the Kalman gain by the error between the observation and the corresponding prior which is the total confirmed cases. The 4 step and the 5 step predictions of each model are not able to reflect the weekly 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 weekly confirmed cases is 2569 and Table 3 lists the predicted peak values of weekly 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 weekly confirmed cases by observing the total confirmed cases.
FIGURE 3. Real-time predictions of total confirmed cases

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

| Model   | Peak/ARE | 1 step     | 2 step     | 3 step     | Average ARE |
|---------|----------|------------|------------|------------|-------------|
| SEIR-SEI | 2582/7.3% | 2054/20% | 1725/32% | 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% |           |

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 root mean square (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 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 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.

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 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 $\beta_H, \beta_M$ 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. 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 $\beta_H, \beta_M$ 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
We would like to thank Dr. Aram Vajdi for providing mosquito population data.

REFERENCES
[1] M. G. Guzman, “A new moment for facing dengue?,” Pathogens and global health, vol. 109, no. 1, 2, 2015.

[2] C. Fitzmaurice et al., “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 oncology, vol. 3, no. 4, pp. 524–548, 2017.

[3] S. Bhatt et al., “The global distribution and burden of dengue,” Nature, vol. 496, no. 7446, pp. 504–507, 2013.

[4] J. Rajarethinam et al., “Dengue in Singapore from 2004 to 2016: cyclical epidemic patterns dominated by serotypes 1 and 2,” The American journal of tropical medicine and hygiene, vol. 99, no. 1, p. 204, 2018.

[5] E. J. Luna et al., “A cohort study to assess the incidence of dengue, Brazil, 2014–2018,” Acta tropica, vol. 204, p. 105313, 2020.

[6] W.-J. Chen, “Dengue outbreaks and the geographic distribution of dengue...
FIGURE 8. Real-time predictions for 2020 dengue outbreak in Singapore

FIGURE 9. Real-time predictions 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

V. Viboud et al., “The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt,” Epidemics, vol. 22, pp. 13–21, 2018.

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, 2016.

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,” Scientific reports, vol. 9, no. 1, pp. 1–9, 2019.

R. M. Anderson and R. M. May, Infectious diseases of humans: dynamics and control. Oxford university press, 1992.

Q. Yang et al., “Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China,” medRxiv, 2020.
[15] J. Shaman and A. Karspeck, “Forecasting seasonal outbreaks of influenza,” Proceedings of the National Academy of Sciences, vol. 109, no. 50, pp. 20425–20430, 2012.

[16] L. Esteve and C. Vargas, “Analysis of a dengue disease transmission model,” Mathematical biosciences, vol. 150, no. 2, pp. 131–151, 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, p. e49085, 2012.

[18] M. Otero, D. H. Barmak, C. O. Dorso, H. G. Solari, and M. A. Natiello, “Modeling dengue outbreaks,” Mathematical biosciences, vol. 232, no. 2, pp. 87–95, 2011.

[19] R. A. Erickson, S. M. Presley, L. J. Allen, K. R. Long, and S. B. Cox, “A dengue model with a dynamic Aedes albopictus vector population,” Ecological Modelling, vol. 221, no. 24, pp. 2899–2908, 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, 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,” Scientific reports, vol. 8, no. 1, pp. 1–8, 2018.

[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, p. e1003583, 2014.

[23] R. E. Kalman, “A new approach to linear filtering and prediction problems,” 1960.

[24] G. Evensen, “Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics,” Journal of Geophysical Research: Oceans, vol. 99, no. C5, pp. 10143–10162, 1994.

[25] J. P. Jensen, “Ensemble Kalman filtering for state and parameter estimation on a reservoir model,” Master’s Thesis, Institutt for teknisk kybernetikk, 2007.

[26] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser, “Dual state–parameter estimation of hydrological models using ensemble Kalman filter,” Advances in water resources, vol. 28, no. 2, pp. 135–147, 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 2015 IEEE Power & Energy Society General Meeting, 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 Negl Trop Dis, vol. 9, no. 8, p. e0004002, 2015.

[29] Yang, H. M., M. L. G. Macoris, K. C. Galvani, M. T. M. Andrighetti, and D. M. V. Wanderley. “Assessing the effects of temperature on dengue transmission,” Epidemiology & Infection 137, no. 8 (2009): 1179-1187.

[30] L. Esteva and C. Vargas, “Analysis of a dengue disease transmission model,” Mathematical biosciences, vol. 232, no. 2, pp. 87–95. 2011.

[31] CHUNLIN YI received the B.S. degree in electrical engineering from the Harbin institute of Technology, Harbin, China in 2018. He started his Ph.D. study in electrical and computer engineering under the supervision of Dr. Caterina M. Scoglio in Kansas State University from 2019. His research interests include network science, modeling of epidemics and Kalman filtering.

[32] LEE W. COHNSTAEDT received his PhD 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.

[33] CATERINA M. SCOGLIO (Senior Member, IEEE) received the Dr. Eng. degree from the “Sapienza” Rome University, Italy, in 1987. She is the Paslay Chair Professor of Electrical and Computer Engineering with Kansas State University. Her main research interests include the field of network science and engineering. Before joining Kansas State University, she was with the Fondazione Ugo Bordoni from 1987 to 2000, and Georgia Institute of Technology from 2000 to 2005.