scientific reports

Check for updates

OPEN

Incorporating human mobility data improves forecasts of Dengue fever in Thailand

Mathew V. Kiang[1,12], Mauricio Santillana[2,3,12], Jarvis T. Chen[4], Jukka-Pekka Onnela[5], Nancy Krieger[4], Kenth Engø-Monsen[6], Nattwut Ekapirat[7], Darin Areechokchai[8], Preecha Prempree[8], Richard J. Maude[7,9,11,13] & Caroline O. Buckee[10,11,13]✉

Over 390 million people worldwide are infected with dengue fever each year. In the absence of an effective vaccine for general use, national control programs must rely on hospital readiness and targeted vector control to prepare for epidemics, so accurate forecasting remains an important goal. Many dengue forecasting approaches have used environmental data linked to mosquito ecology to predict when epidemics will occur, but these have had mixed results. Conversely, human mobility, an important driver in the spatial spread of infection, is often ignored. Here we compare time-series forecasts of dengue fever in Thailand, integrating epidemiological data with mobility models generated from mobile phone data. We show that geographically-distant provinces strongly connected by human travel have more highly correlated dengue incidence than weakly connected provinces of the same distance, and that incorporating mobility data improves traditional time-series forecasting approaches. Notably, no single model or class of model always outperformed others. We propose an adaptive, mosaic forecasting approach for early warning systems.

More than half the world's population is at risk of infection from the dengue virus, which causes an estimated 390 million infections[1] and 25,000 deaths per year[2,3]. The dengue pathogen is spread in urban and peri-urban areas by invasive mosquitoes belonging to the Aedes complex. As a result, dengue has emerged as a major threat in the context of a rapidly urbanizing, globally connected world[3–5]. For example, despite the general decline in the incidence of other communicable diseases, the incidence of dengue fever has doubled every 10 years since 1990[6]. The rapid geographic expansion of the vector suggests there will be a continuing emergence of dengue globally[3–5]. Currently, there is no drug treatment for dengue[7,8] and only a partially effective vaccine, which cannot be used in seronegative individuals[9]. Therefore, despite the mixed results of vector control efforts[8], targeted and thorough vector control approaches, hospital readiness, and risk communication can improve public health preparedness for seasonal outbreaks. Fundamental to the success of these preparations is data on the burden of disease in different areas, and some sense of how an epidemic may progress in the near term and on local spatial scales relevant to national control programs.

Forecasting the epidemic trajectory of dengue on weekly or monthly timescales remains a relatively new science for infectious diseases[10–23]. Unlike weather and climate forecasting, where physical laws dictate the dynamics of the system, the social and biological dynamics that drive infectious disease outbreaks make forecasting dengue epidemics challenging. Recurring epidemics, as opposed to novel pathogens emerging for the first time, occur against a backdrop of shifting population immunity, which is difficult to quantify. Complicating surveillance, pathogens like dengue are primarily reported based on symptoms rather than laboratory confirmation. Like

[1]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA. [2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [3]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [4]Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Telenor Research, Oslo, Norway. [7]Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand. [8]Bureau of Vector Borne Disease, Ministry of Public Health, Nonthaburi, Thailand. [9]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [10]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [11]Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, 5th Floor, Boston, MA 02115, USA. [12]These authors contributed equally: Mathew V. Kiang and Mauricio Santillana. [13]These authors jointly supervised this work: Richard J. Maude and Caroline O.Buckee. ✉email: cbuckee@hsph.harvard.edu

influenza and malaria, dengue causes non-specific symptoms, fever in particular, so reporting reliability and time lags impact data quality[24–26]. Despite these complexities, routine forecasting is an important priority for national dengue control programs[8,11].

There has been a recent surge of interest and success in building forecasting models for seasonal epidemics of dengue fever[10–21]. A distinction can be made between mechanistic epidemiological models and statistical models. In mechanistic models, the mode of transmission (in this case, mosquito-borne and strong temperature dependence) is built into the model and drives the predicted infection dynamics. In contrast, statistical models rely on the identification of past epidemiological activity patterns and historical correlations with external data streams, often generated by human behavior on Internet search engines or social media, to monitor disease activity and predict future outbreaks. Mechanistic models aim at providing biological insight and a basis for interpretation, but for socially and environmentally complex infections like dengue, these models are often challenging to parameterize. Dengue is particularly challenging as it is composed of multiple immunologically distinct strains and relies on the interaction of mosquito and human population dynamics and microclimate variability. Metapopulation models have been developed to incorporate the spatial dynamics of dengue outbreaks, modeling each area with a set of location-specific parameters and linking the areas through estimated migration of individuals. Metapopulation models play in an important role in our understanding of epidemic outbreaks across spatial regions[27–29], synchronicity between regions[30], oscillations of epidemics[31], and strategies to reduce transmission[32]. Despite their importance in understanding dynamics, mechanistic models, and metapopulation models in particular, may lack sufficient data for appropriate parameterization, and are often not feasible in a forecasting context. As a result, statistical models have been more successful for outbreak preparedness for which the modeling goal is to provide quantitative, relatively short-term predictions with explicit uncertainty[10,12–21,27–29].

Most statistical forecasting approaches for dengue have been based on autocorrelation in case data, often incorporating environmental information due to the importance of temperature and other factors to the availability of mosquitoes and variation of the incubation period of the virus in the vector. Many of these have focused on long-term predictions of dengue at the city level[15,21,33,34], or larger regions within a specific country[14,16,17,19]. Models often show mixed success with high prediction accuracy in the immediate forecasting horizons (e.g., 1–2 months) and rapid decay at longer time horizons (e.g. 3–6 months). It is unclear if weather or climate variables substantially improve forecasting; at least one study that systematically looked at different model parameters for autoregressive models, with and without a wide range of climate variables, across states in Mexico found no conclusive improvement[12]. More recently, ensemble models have become a powerful way to combine different approaches in order to leverage the strengths of each while minimizing the weaknesses[23,35]. This approach has recently been applied to dengue[13]. Others have incorporated new sources of data from internet search terms to predict dengue nationally[18], employed novel statistical methods to predict dengue in San Juan, Puerto Rico[36], or combined common climate covariates with generalized additive models to predict annual incidence of dengue hemorrhagic fever[10].

Although dengue outbreaks among new human populations, both across long distances[37,38] and within local communities[39], is spread primarily via human mobility[5], incorporating this aspect of the spatial connectivity between locations within forecasting frameworks has been challenging. Current forecasting models, both mechanistic and statistical, either ignore or make crude assumptions about how populations are connected by travel. Parameterizing human mobility is challenging due to a paucity of relevant data streams, particularly in low-income settings. We have previously used mobile phone records to quantify national movements and showed that they provide improved prediction for dengue outbreaks in Pakistan[5]. Specifically, we used a gravity model to parametrize human mobility in a mechanistic framework because dengue was emerging into naïve populations, where statistical methods could not be used. Others have used daily commuting data to model mobility using a radiation model, which in turn is used to parameterize a mechanistic model[40]. Although considerable difficulty remains in accessing mobile phone records or other scalable data sources about mobility, it is clear that gravity models, radiation models, and other proxies for travel measures may perform poorly in many settings[41].

To date, almost all efforts to forecast dengue have either focused on optimizing a single modeling framework across regions, fitting parameters individually, or analyzed multiple models for a particular location. Few statistical models used for forecasting dengue incorporate spatial dependencies and none incorporate information about mobility patterns. Here, we contribute to the existing literature by using 7 years of monthly dengue data (2010–2016) from Thailand, which has a developed dengue surveillance program, and mobility data from approximately 11 million mobile phone subscribers to show that long-distance provinces that are more strongly connected by human mobility have more highly correlated dengue incidence than weakly connected provinces. We compare model structures incorporating time-series approaches or spatial dependencies, and mobility data, finding that this improves model prediction, but no individual approach provides the best performing model in all locations over all time horizons. We quantify the error for each province in Thailand, showing that provinces in the north of the country are more difficult to forecast with confidence than those in the south, regardless of model choice, and that different models' performances may be linked to demographic and social factors such as population density and gross provincial product per capita. We propose that mosaic forecasting approaches, which dynamically adapt over time and space, and end up using the best model for that location and time period, are likely to be the most effective for use in early warning systems in national control programs.

Results

No one-size-fits-all: forecasting performance varies in space and time. We compared several forecasting approaches for the 77 Thai provinces to assess how model performance varied by region and over time, and to measure the impact of integrating the mobility data. Specifically, for each province, we fit four models: (1) local (non-spatially dependent) models commonly used for dengue; specifically, seasonal autore-

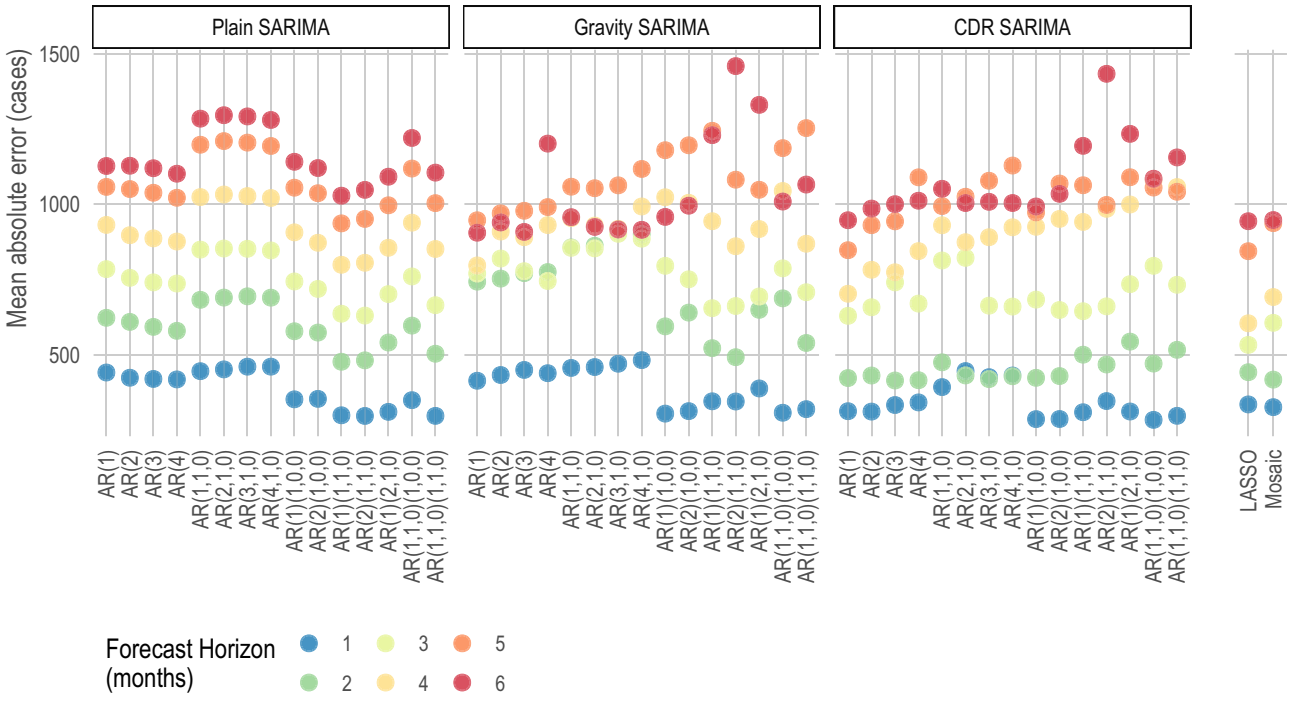

Figure 1. Mean absolute error (MAE) for all Bangkok models. The mean absolute error (y-axis) expressed as number of cases for each model (x-axis) and for each forecast. Models are grouped as SARIMA with no exogenous variables (Plain), SARIMAs with the top 5 most connected regions based on the predicted trips from a gravity model (Gravity SARIMA), and SARIMAs with the top 5 most connected regions based on CDR data (CDR SARIMA). The rightmost models show a data-driven network model, denoted as LASSO, since it is based on a least absolute shrinkage and selection operator prediction model, and mosaic model.

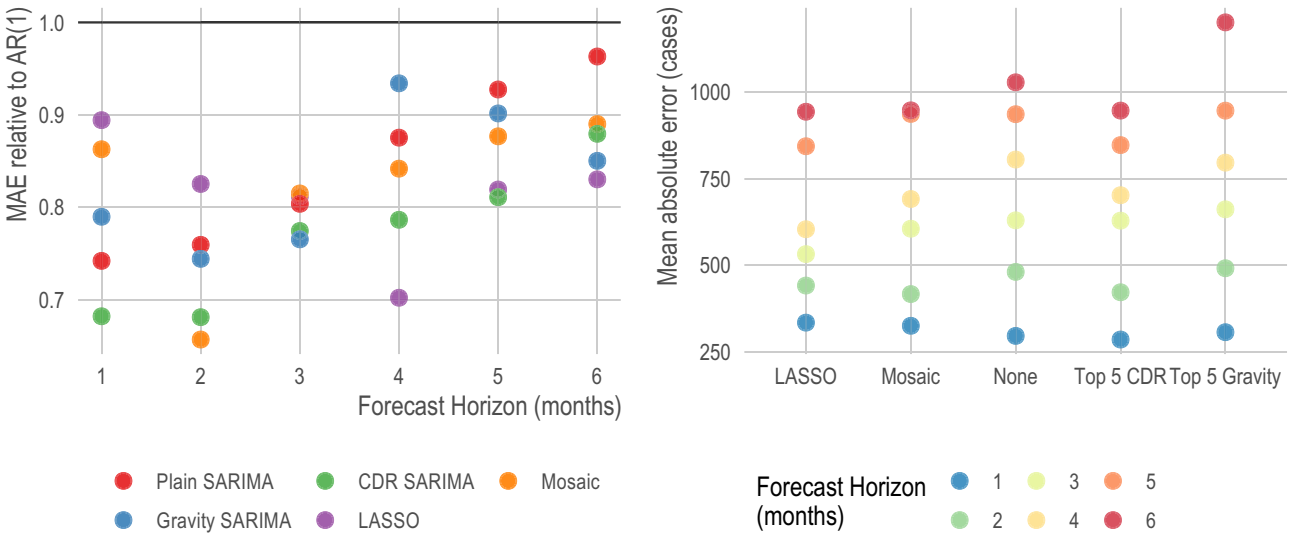

Figure 2. Comparing the best models for Bangkok, by model type. Focusing only on the best performing model for each model type and each time horizon, we show the relative mean absolute error (left panel) and the mean absolute error (right panel). On the left, the baseline of comparison is the traditional AR(1) model and the y-axis can be interpreted as the improvement over this baseline—i.e., a value of .9 indicates a 10% improvement. We show that both the Plain SARIMA (red) and CDR SARIMA (green) models perform better than the LASSO model at earlier forecasting horizons but perform worse at later horizons.

gressive integrated moving average models (Plain SARIMA) across a grid of parameters, (2) SARIMA models that use information from the top five most connected provinces (in terms of number of incoming trips) based on mobile phone data (CDR SARIMA), (3) SARIMA models that use information from the top five most connected provinces (in terms of predicted number of incoming trips) based on our gravity model estimates (Grav-

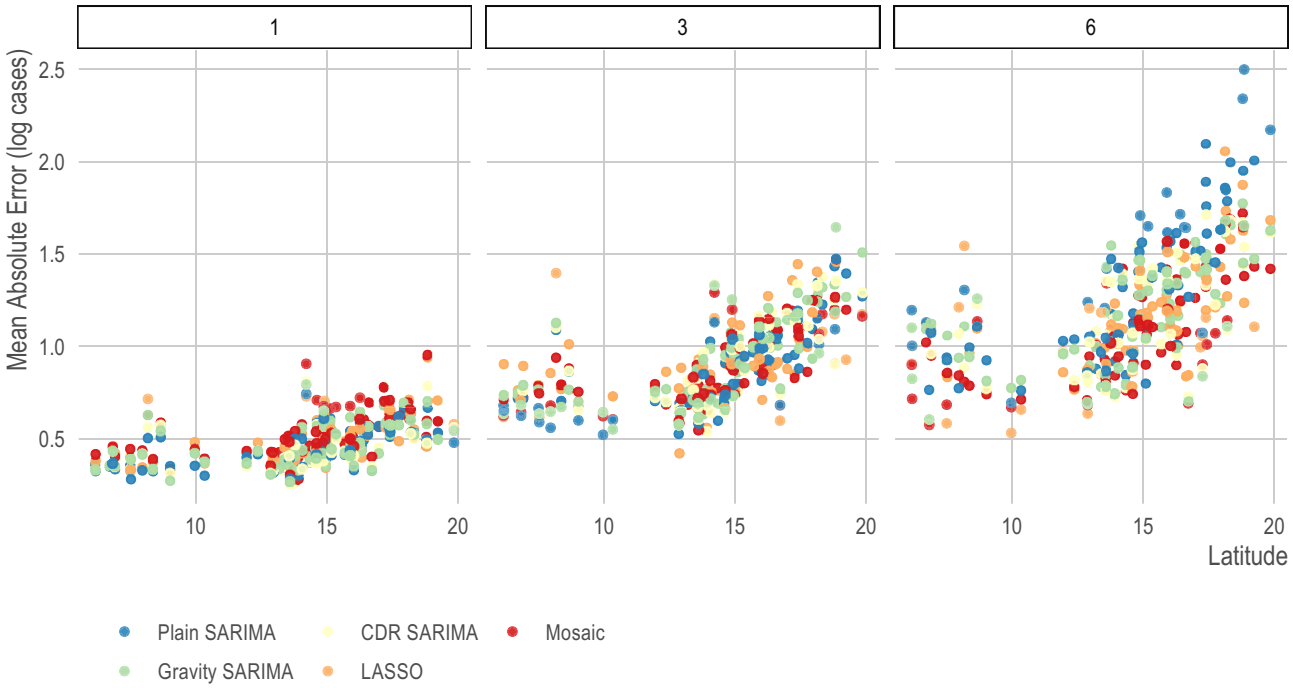

Figure 3. Mean absolute error for the best model in each class at t + 1, t + 3, and t + 6 forecasting horizons for all provinces. The mean absolute error (y-axis) on the prediction (i.e., log) scale of the best model for each class for all provinces (x-axis). Provinces are ordered by latitude (x-axis, right is more northerly). There is a general decline in predictive power at farther forecasting horizons and at more northerly provinces; however, no single model or class of model performs best across all areas and all prediction horizons.

ity SARIMA), and (4) a data-driven network approach, based on a regularized regression approach, that predicts dengue incidence in a given location potentially using dengue incidence from every other location as input (LASSO; see "Materials and methods" and reference[49] for details).

Figure 1 illustrates the results of all models at all forecasting horizons for Bangkok (see SI Appendix, Text S1 for online-only results for all other provinces). At early forecasting horizons (i.e., 1-month and up to 3-months ahead time horizons), all models performed well, with the CDR SARIMA and Gravity SARIMA models outperforming the Plain SARIMA models by about 5–10% (Fig. 2) as captured by the mean absolute error. After the 3-month forecasting horizon, the Plain SARIMA model performance drops substantially faster than all other models. Importantly, the grouping of out-of-sample prediction errors, across forecasting horizons, tended to be closer in the LASSO models, indicating that across forecasting horizons, the network models lose predictive power more slowly than the SARIMA-based models. We present all plots for all provinces in an online repository (SI Appendix, Text S1).

Across other provinces, the observations of model performance for Bangkok are similar. Specifically, all models performing well in the near time, Gravity and CDR SARIMA models usually outperform Plain SARIMA models, and there is lower variation in prediction error when using the LASSO models. Importantly, no single model or class of model outperformed others across all provinces or all forecasting horizons (Fig. 3; SI Appendix, Fig. S5). We found that across all model types, provinces in the south of the country had lower prediction errors compared to those in the north of the country (Fig. 3). This difference in forecasting power was particularly pronounced on farther forecasting horizons. For example, when comparing the out-of-sample prediction errors of the CDR SARIMA to the Plain SARIMA, the CDR SARIMAs were worse in 8 tasks for forecasting horizons of 1–3 months and better in only 3 tasks with no statistically significant difference in the remaining 220 prediction tasks. However, for forecasting horizons of 4–6 months, the CDR SARIMA outperformed the Plain SARIMA in 40 tasks and only underperformed in 8 with no statistically significant difference in the remaining 183 tasks (SI Appendix, Fig. S6).

We measured the characteristics of provinces in which different models performed better or worse and found that the Plain SARIMA models performed similarly when comparing top and bottom deciles of total number of dengue cases, median number of monthly dengue cases, median monthly rate of dengue, population density, and GPP per capita. In contrast, the LASSO and mobility-augmented SARIMA models performed better in places with higher total annual cases, higher population, and lower GPP per capita (see SI Appendix, Fig. S8–S13), suggesting systematic and generalizable differences in model performance that — with more validation and in combination with geographic variation in model performance—could be used to inform model choice.

We show the feasibility of combining different classes of models by using a simple winner-takes-all voting system approach we named an adaptive mosaic model. This ensemble model selects the best performing model for each province and forecasting horizon based on the out-of-sample prediction error of previous 3 months, which allows the underlying base model to change over time (Fig. 4). When comparing the out-of-sample

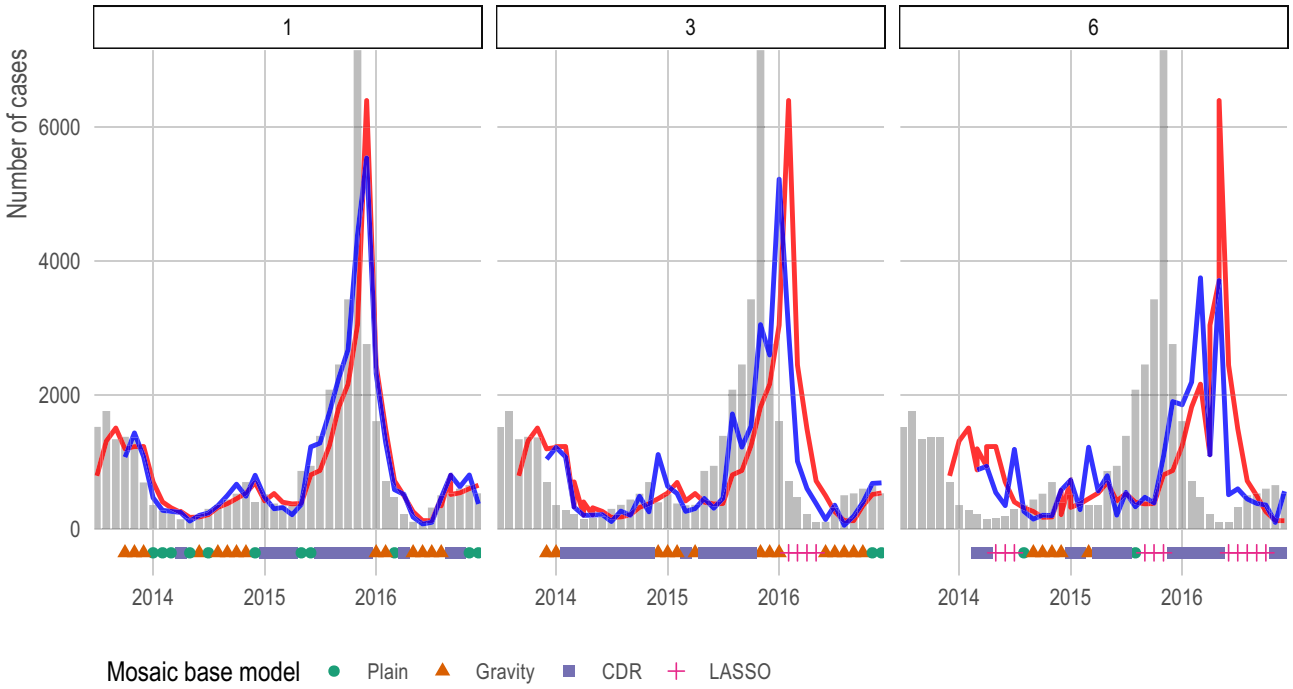

Figure 4. Mosaic model vs AR(1) for Bangkok at t + 1, t + 3, and t + 6 forecasting horizons. We show the predictions for a simple mosaic model at t + 1, t + 3, and t + 6 forecasting horizons for Bangkok in blue. For comparison, we show predictions from an AR(1) in red and observed cases in grey bars. Under each bar, we indicate the base model selected by the mosaic ensemble using a winter-takes-all approach based on the previous three out-of-sample prediction months. The t + 6 forecasting horizon presented a significant challenge for all models, but the mosaic model adapted more quickly and did not over-predict relative to the AR(1).

prediction errors to an AR(1) model, the mosaic model outperforms the AR(1) in 107 tasks (i.e., province and forecasting horizon), underperforms the AR(1) in 3 tasks, and is not statistically significantly different from an AR(1) in the remaining 352 tasks (SI Appendix, Fig. S7). At the 6 month forecasting horizon, a difficult prediction task for any model, we note that no models were able to predict the incidence peak in 2015; however, the adaptive mosaic model compensated more quickly and did not overshoot its prediction relative to the AR(1) model (Fig. 4). Further exploration of location-specific and task-specific voting predictions systems is outside of the scope of this study but should be explored in future research efforts.

Gravity models under-predict long-distance travel to and from Bangkok. To assess the role of inter-province migration, we analyzed the call data records (CDR) of approximately 11 million mobile phone subscribers between August 1, 2017 and October 19, 2017. At the time of data collection, the mobile phone operator had about 26% of the market share and was the third largest provider in Thailand. Since travel patterns remained stable over our period of observation (coefficient of variation: 1.3%; SI Appendix, Fig. S1), we calculated average daily journeys between all pairs of provinces in both directions, and compared observed mobility in the CDR data to expected mobility based on gravity models (see "Materials and methods") assuming travel over our time period is consistent with travel for the rest of the year (SI Appendix, Fig. S2). We found that the routes of travel that deviate significantly from gravity model-based predictions in both directions are focused on Bangkok (Fig. 5), with more travel than expected from long distances around the country such as Phuket and Bangkok itself (Fig. 5, left), and less travel than expected within and around the city (Fig. 5, right). These hot and cold spots, where higher or lower than expected travel was observed, were robust to the gravity model coefficients used (SI Appendix, Table S1).

At longer distances, strongly connected provinces show higher correlation in dengue incidence than weakly connected provinces. In Thailand, dengue follows a seasonal cycle across all 77 provinces (Fig. 6), with variation in the timing of onset and epidemic peak in different locations over our period of observation[42]. We analyzed the correlation between clinical cases in each province with different time lags between all pairs of provinces. Figure 7 shows the relationship between the correlation in dengue cases between pairs of provinces, stratified with respect to geographic distance and mobility measured using mobile phone data. Consistent with previous studies[43–46], the correlation of dengue incidence between provinces is strongest when they are close to each other and declines with distance and over time (i.e. the 3-month lagged correlation is weaker than the 1-month lagged correlation). For provinces less than 1,000 km apart, human mobility estimated using mobile phone data does not appear to impact the correlation of clinical cases. For longer distances, however, more highly connected locations show higher correlation in clinical dengue cases than locations the same

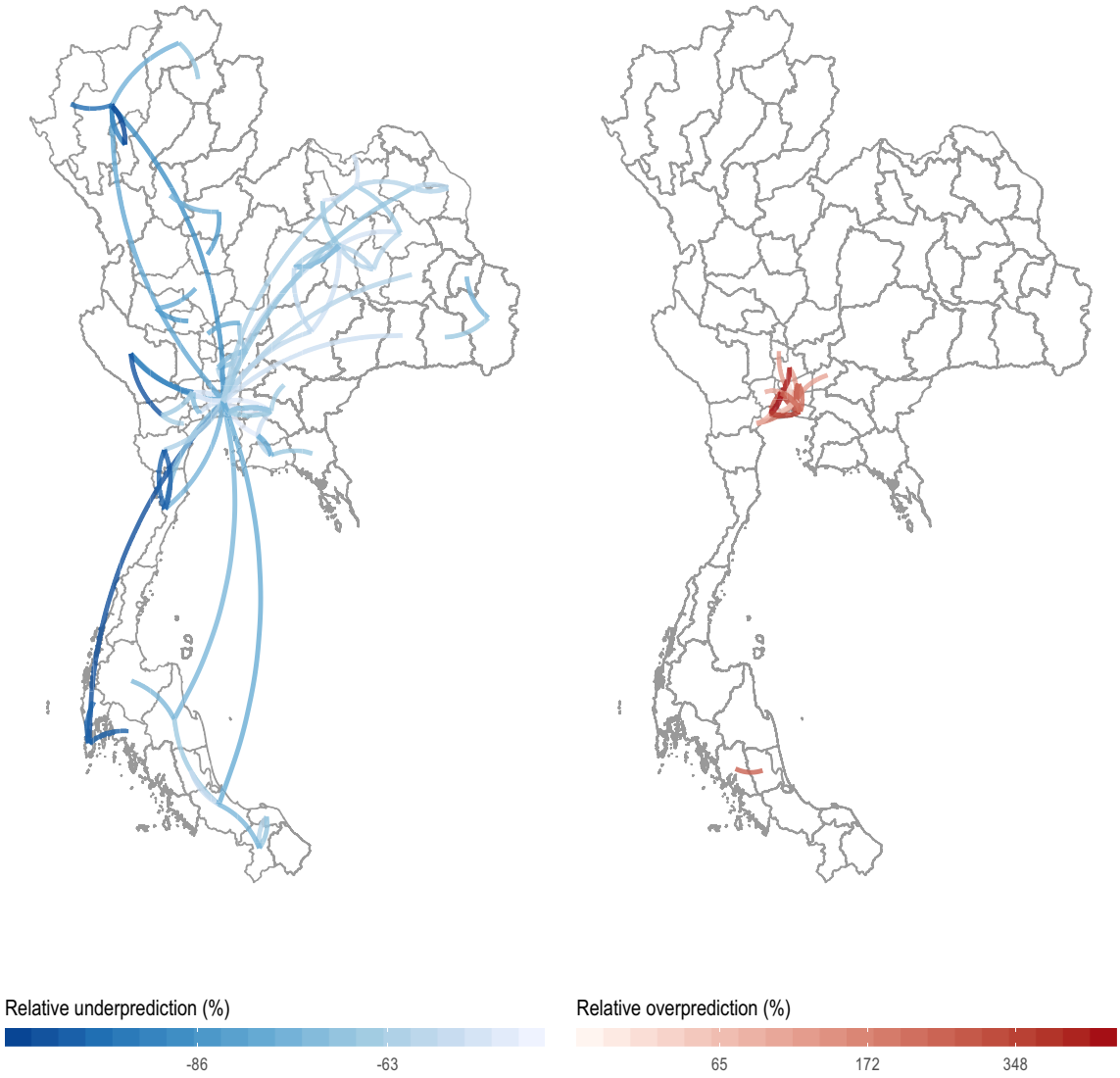

Figure 5. Under- and over-prediction of outlier travel. Relative under-prediction (left) and over-prediction (right) comparing observed mobility data (from CDRs) to estimated mobility data from the best fit gravity model. We defined relative prediction error as 100%*(PredictedTrips − ObservedTrips)/ObservedTrips. We highlight only observations with Cook's distance greater than five times the average Cook's distance. Note that Bangkok (center of the map) is central to much of the over- and under-prediction outliers with most over-prediction near Bangkok. All plots were made using ggplot2[55] in R 4.0.1[56].

distance apart but with low observed connectivity (Fig. 7). Note that some but not all of these long-distance connections are locations with international airports (SI Appendix, Fig. S3), and provinces connected by airports have higher correlation than those that are not connected by airports (SI Appendix, Fig. S4).

Discussion

Dengue forecasting remains an important public health challenge in Thailand and other endemic countries, especially at farther forecasting horizons. Given the complexity of dengue transmission, statistical forecasting approaches like those examined here have been shown to produce meaningful disease estimates in multiple locations and may therefore be suitable for immediate use by national control programs. In addition, we have shown that integrating additional data streams, such as information about human mobility, can improve forecasts in many areas, but the added benefit will be specific to the area and time horizon of interest. The interesting geographic variation in forecasting accuracy, which is not linked to population density or GPP per capita, may reflect the proximity to international borders with countries where frequent migration occurs. Overall, no single modeling approach can be expected to provide an optimal early warning system across all areas, even within a single country or region, or across all time horizons. So adaptive, mosaic forecasts are likely to provide the most effective approach. This type of approach could be easily integrated within the data platforms recently developed in Thailand[11], which are flexible enough to accommodate different modeling approaches and forecasts.

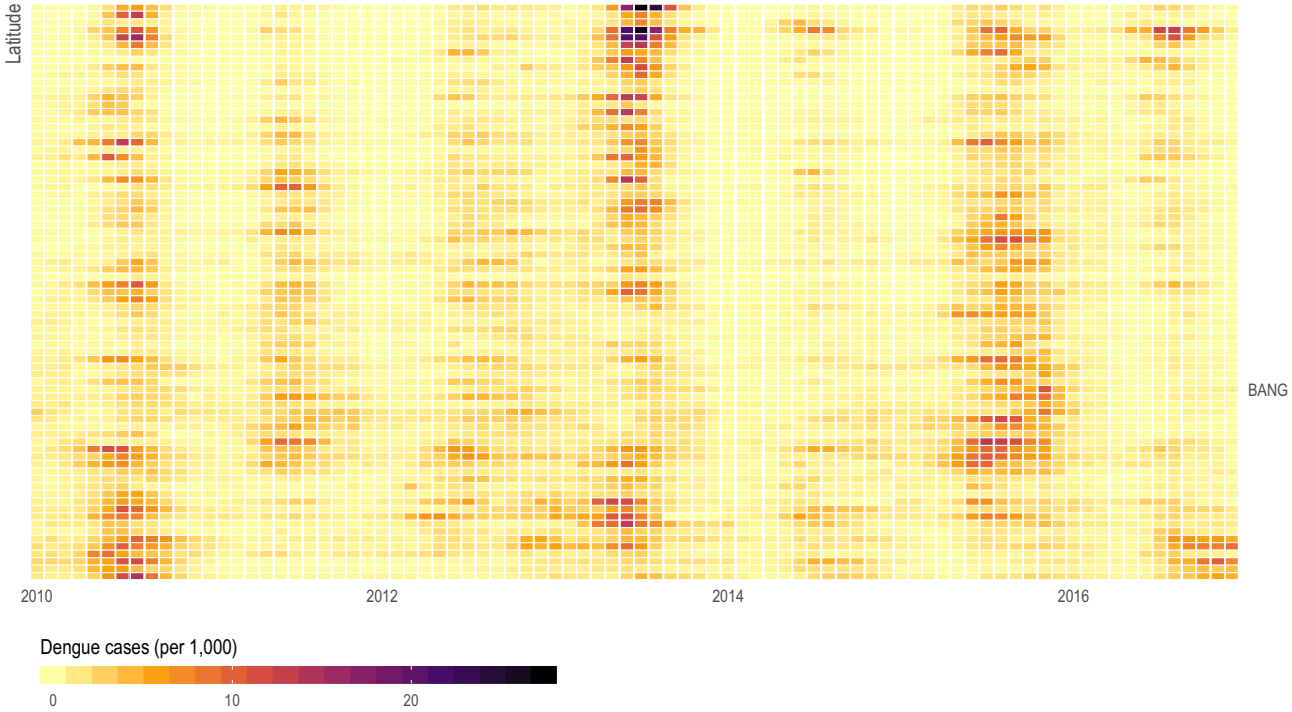

Figure 6. Monthly dengue incidence by province. Monthly crude incidence of dengue (per 1000 person-years) by province (y-axis) ordered by centroid latitude (higher is more northern) over 7 years of observation (x-axis). Dengue in Thailand follows a seasonal cycle with geographic variation in both the timing of onset and peak of the epidemic.

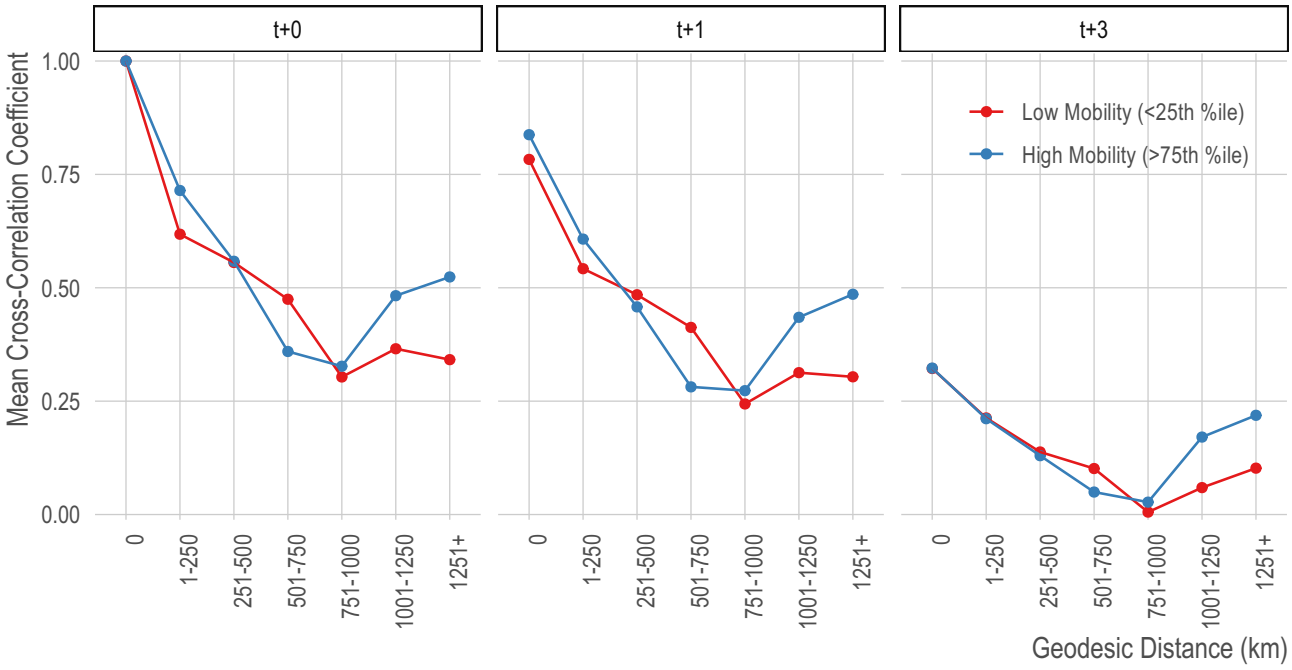

Figure 7. Correlation of province-level dengue by distance, at different time lags. We show the mean cross-correlation coefficient (y-axis) for pairs of provinces at binned distances (x-axis; 0 indicates correlation of an area with itself) for synchronous dengue (left panel) and lagged by 1 month (middle panel) and 3 months (right panel). The lines are separated based on the connectivity of pairs of provinces where the red line shows the bottom quartile of provinces in terms of incoming and outgoing travel and the blue line shows the top quartile. Bangkok, an important travel hub, is in the approximate center of Thailand and between 700 and 800 km from all other provinces, therefore the last two distance categories do not include Bangkok.

We show that simple network methods (that implicitly incorporate human mobility) can improve upon commonly-used local SARIMA models. Also, given that the network-based approach we studied relied only on dengue case count data routinely collected by most endemic countries, we envision that similar approaches may be easily extended, and may prove to be meaningful, in many other locations around the Globe. The regularized multi-variate regression framework can also flexibly identify and incorporate additional province-level data, time lags, and other factors in the predictive model, that could be used as a hypothesis-generation tool that may capture temporal changes in inter-regional human mobility. We highlight the fact that even though the mobility data we used covered only a small fraction of time represented in the dengue case data (3.2%; i.e., 81 days vs 7 years), it was still able to improve the local (non-augmented) SARIMA, suggesting that even relatively coarse travel information would improve naïve SARIMA models. Although mobile phone data is challenging to obtain, the coarse granularity of mobility information that we used completely protects individual subscriber privacy while adding substantially to forecasting performance. Since it is continuously collected, there is no reason these data could not be aggregated by mobile operators and provided on a relatively frequent basis to disease control programs. A limitation of using CDR to model dengue transmission is that it reflects movement patterns of the entire population whereas dengue tends to occur more in children and young adults in urban areas[42].

As governments prioritize how and where to spend money to improve dengue surveillance, our study suggests new regularized regression frameworks that incorporate mobility data can improve forecasts substantially. Any forecasting model will depend on the quality of the case data that it is trained upon, highlighting the primary importance of good epidemiological data. A limitation of this work is that most dengue cases in Thailand, as in most countries, are not confirmed with a diagnostic test, instead relying in syndromic surveillance. This can be unreliable with the case definition for dengue fever overlapping substantially with other causes of acute febrile illness and the completeness of the data relying on individual healthcare workers to complete the reporting forms. Thus, much of the money for better dengue forecasting should be focused on faster and better dengue case detection, more widespread diagnostic testing, sentinel surveillance of serotypes, a robust computational framework for sharing case data across regions to be analyzed centrally, and capacity building within control programs. In addition, we note that dengue in Thailand follows a cyclical, multi-year pattern of higher incidence[44], which is not fully captured in our observation window of 7 years. In addition to better quality data, more historical data will be necessary for improving forecasting models that incorporate these longer period cycles.

We are limited to the use of call detail records over the course of 81 days in a single year and must assume that the relative mobility in this period is representative of the full 7 years of clinical case data. Due to legal, regulatory, and logistical reasons, longer historical mobility patterns are often not feasible. For example, mobile operators do not store this information, and are often not allowed to store this information, for more than a few months. Previous research, using other sources of data for human mobility, suggests holiday and seasonal fluctuations in mobility affect the relative routes of travel within countries but that within-country mobility is remarkably stable[47]. Incorporating seasonal and holiday fluctuations in model predictions is an important area for future research. Similarly, our mobility data are memory-less and intermediate locations between two provinces are not recorded yet may play an important role in transmission dynamics. Additional mobility data, perhaps outside the regulatory constraints of mobile phone operators, is necessary to assess this possibility.

Materials and methods

Dengue incidence data. We obtained monthly dengue case counts for over 7,000 subdistricts in Thailand from the Ministry of Public Health. These data are not available publicly and are used with the permission of the Ministry of Public Health. They consist of monthly dengue incidence counts from January 2010 through December 2016, by mutually-exclusive disease type (i.e., dengue fever, dengue shock syndrome, or dengue hemorrhagic fever). We aggregated these data to the province level and overall dengue case counts. In our data, there was a national average of 91,000 dengue cases per year with a range of 39,368 (2014) to 145,600 (2013) cases per year.

Mobile phone data. To assess inter-province travel, we analyzed call data records (CDRs) of approximately 11 million mobile phone subscribers between August 1, 2017 and October 19, 2017. At the time of data collection, the mobile phone operator had about 26% of the market share and was the third largest provider in Thailand. In order to ensure the privacy of the mobile phone subscribers, and in compliance with national laws and the privacy policy of the Telenor group, special considerations were taken with the CDRs. First, only the mobile operator had access to the CDR and all data processing was performed on a server owned by, and only available to, the operator, thus ensuring that detailed records never left the operator or Thailand. Second, the operator provided researchers with a list of approximate cellular tower locations. For every tower location, we returned a corresponding, unlinked geographic identifier ("geocode") of the nearest subdistrict. Mobile operator employees then aggregated the detailed CDRs up to the researcher-provided geocodes. Further spatial and temporal aggregation was performed by the researchers. These data are not publicly available and are used with the permission of Telenor Research.

To quantify travel, every subscriber was assigned a daily "home" location based on their most frequently used geocode. We tabulated daily travel between a subscriber's home location on one day relative to the day before. Trips were aggregated to geocode-to-geocode pairs for every day and thus are memoryless — preventing the ability to trace a user (or group of users) across more than two days or more than two areas. We normalized the number of trips from geocode i to geocode j by the number of subscribers at geocode i. We then multiplied this proportion by the estimated population at geocode i to get the flow from i to j. This assumes that subscribers are more or less uniformly distributed across provinces (weighted by the population in each province). While this assumption cannot be fully tested, there is a strong correlation (Pearson's $r = 0.90$) between subscribers and population for each province.

On average, 11.4 million subscribers (16.7% of the total population) recorded at least one event (i.e., phone call, text message, internet activity) per day (SI Appendix, Fig. S1). At both the national and provincial levels, no significant deviations from the number of subscribers or the numbers of trips occurred during this time period. For example, at the national level, the coefficient of variation for daily number of subscribers was 1.3%. Therefore, we used the mean number of trips over this time period as our estimate of inter-province travel.

Population, gross provincial product per capita, and distance estimates. To estimate province-level population, we used the United Nations-adjusted 2015 population estimates from WorldPop[48], which combines remote-sensing data with other data sources to create random-forest-generated population maps. Each file contains the estimated population per pixel and was overlaid with the official administrative shapefile. We then summed the value of all pixels within each province. We used publicly available 2015 gross provincial product per capita provided by the Office of the National Economic and Social Development Board of Thailand[49]. The concept of "distance" is flexible in the gravity model and geodesic distance often ignores important geographical (e.g., mountain ranges) or social and behavioral constants to human mobility. In addition to calculating geodesic distance between provinces, we calculated road distance and travel time based on OpenStreetMap data using Open Street Routing Machine[50].

Comparing observed and predicted travel. We compared observed travel between provinces with CDRs to those estimated by a gravity model with three different measures of distance: geodesic distance, road distance, and travel time. The gravity model is a popular econometric model[51], often used to estimate mobility between areas[52]. The basic gravity model is:

$$Y_{ij} = k\frac{P_i^{\alpha}P_j^{\beta}}{D_{ij}^{\gamma}}$$

where Y_{ij} be the number of people who move from area i to area j, k is a constant term, P_i is the population in area i, P_j is the population in area j, and D_{ij} is some measure of distance between i and j, noting that distance may not be symmetric. The parameters k, α, β, and γ are estimated by fitting a Poisson model:

$$\log\left(\mathrm{Y_{ij}}\right) = \mathrm{k} + \alpha\log(\mathrm{P_i}) + \beta\log\left(\mathrm{P_j}\right) - \gamma\log(\mathrm{D_{ij}}).$$

In addition to the naïve gravity model, we also adjusted for gross provincial product per capita. The best fit according to in-sample error metrics was the adjusted travel time model (SI Appendix, Table S1). We identified outlier observations as those observations with a Cook's distance greater than five times the mean Cook's distance.

Quantitative methods. We evaluated the predictive accuracy of two different types of models: (1) one data-driven network approach built using an L_1-regularized regression approach (the least absolute shrinkage and selection operator, LASSO) and (2) autoregressive integrated moving average (ARIMA) models both with and without a seasonal component (SARIMA). In addition, for the mobility-augmented autoregressive models, human mobility is accounted for by also including lagged case data from the top five areas (i.e., origins) of travelers as covariates in the model. We compared both sets of autoregressive models to the network approach predictions using a sliding window of observation and rolling forecast target as described below.

Network models. Based on a previous model designed to leverage spatially-correlated cases of influenza[53], we fit a multivariate linear regression on the log of dengue case counts for area i in month t with the log of dengue case counts in areas j at time $t-h$ where h is our forecasting horizon as the covariates. Let $y_{i,t} = \ln(\mathrm{c_{i,t}} + 1)$ where c is the count of cases of location i at time t:

$$y_{i,t} = \beta_{0i} + \sum_{j=1}^{J}\beta_j y_{j,t-h} + \epsilon.$$

We used a sliding window of 42 months and h between 1 and 6. All values of $y_{i,t}$ were standardized to be mean-centered with unit variance in order to ensure the coefficients are not scale-dependent. For all prediction months, there were more areas, 77, or input variables, than observations, 42, and thus this formulation cannot be solved using an ordinary least squares (OLS) approach. To address this, we used an L_1 regularization approach to identify a parsimonious model that uses fewer variables as input than the number of available observations. This penalization approach acts to both prevent overfitting as well as selecting the most informative covariates (i.e., provinces). Specifically, we used the least absolute shrinkage and selection operator, LASSO, which minimizes the same objective function as a regular OLS while penalizing the number of non-zero coefficients with a hyper-parameter λ:

$$\min\left\{\frac{1}{\mathrm{N}}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda\|\beta\|_1\right\}$$

where the magnitude of the hyper-parameter λ is identified using cross validation on the training set. This approach shrinks the coefficients of non-informative or redundant areas to zero and provides for straightforward interpretation of the results allowing for identification of which areas contributed the most predictive power for any given window of observation and target area.

Autoregressive models. As a baseline for comparing model predictions, we used autoregressive integrated moving average (ARIMA) models, which is a common time series method applied to epidemiological modeling and dengue forecasting. These models have been used extensively in dengue prediction efforts and often incorporate a seasonal component called Seasonal ARIMA or SARIMA. Using the $(p, d, q)(P, D, Q)_s$ convention where p indicates the autoregressive order, d indicates the amount of differencing, and q indicates the order of the moving average. The seasonal component, $(P, Q, D)_s$, represent the same parameters with a seasonal period of s months. Additional exogenous variables (i.e., timeseries) can be added as covariates in this framework.

We reduced the parameter space of the SARIMA models using previous literature[12] and our expert opinion. Specifically, we systematically search models with lags of up to 4 months (p = 1, 2, 3, or 4) or 3 years (P = 1, 2, or 3) and include a differencing order up to 1 (d and D = 0 or 1) and exclude all moving averages (q and Q fixed at 0). This results in a set of 15 model parameterizations: eight non-seasonal ARIMAs and seven seasonal ARIMAs. For each parameterization, we perform a univariate SARIMA as well as a mobility-augmented SARIMA. The mobility-augmented SARIMA incorporates the timeseries of cases from the top five connected areas, based on observed mobility, as exogenous covariates. Similar to the LASSO, we used a sliding window of 42 months, and in the case of augmented SARIMA models, we lagged the exogenous covariates by h.

Adaptive mosaic model. We show the feasibility of combining different classes of the above models by using an ensemble approach we call the "adaptive mosaic model." For each province and forecasting horizon, we select the best performing model using a winner-takes-all approach based on the out-of-sample prediction error of the previous 3 months. By repeating this procedure for every prediction month, forecasting horizon, and province, the underlying base model can adapt over time (Fig. 7).

Accuracy metrics and model comparison. Consistent with previous research[10,54], when assessing predictive performance of a single model, we used mean absolute error (MAE) and when assessing the relative performance of two models, we used relative mean absolute error (relMAE). The MAE of the log transformed counts is as follows:

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |\ln(y_t + 1) - \ln(\widehat{y}_t + 1)|$$

where y_t and $\widehat{y_t}$ are the observed and average counts for prediction month t. One strength of this approach is that the MAE will be the same regardless of magnitude as long as the ratios are the same (i.e., 100 and 110 for predicted and observed will result in 1.1, just as 10 and 11 or 11 and 10). This is an important feature given the differences in population size and case counts between provinces.

When comparing model A to model B at forecast horizon h, we take the ratio of their MAEs:

$$relMAE_{A,B,h} = \frac{MAE_{A,h}}{MAE_{B,h}}.$$

To assess the predictive performance of each model, we used retrospective out-of-sample estimates of the mean absolute error, assuming we only had data prior to the time of estimation and based on a 42-month sliding window of observation, such that all models are fit on 42-months of observation and evaluated on the out-of-sample forecast as the model slides forward through the remaining available data. For example, the 6-month prediction for June for 1 year would only include data up to December for the year before and only as far back as 42 months from that December. Since there are 7 years of data and we use half (42 months) to fit the model, this provides an additional 42 months to evaluate prediction error as the window of observation slides forward (noting that the number of months available in the evaluation period is also a function of the prediction horizon). To compare across multiple models (e.g., to find the model with the best t + 1 month forecast in a single province), we used the baseline AR(1) (i.e., ARIMA(1,0,0)(0,0,0) with no exogenous variables) as our reference model. Thus, the relMAE can be interpreted as the relative under- or over-performance of our model compared to a standard epidemiological model, averaged over all prediction months.

To assess the utility of call detail records, for each province and forecasting horizon we selected the best performing model of each class. We then compared the CDR SARIMA to each other class using a Wilcoxon signed-rank test to compare the out-of-sample prediction errors. Statistically significant differences are shown in the province-specific reports (SI Appendix, Text S1) and in Figure S7. Similarly, we compared the proposed mosaic model to a simple AR(1) using a Wilcoxon signed-rank test (SI Appendix, Fig. S8).

Received: 8 June 2020; Accepted: 19 November 2020
Published online: 13 January 2021

References

1. Bhatt, S. *et al.* The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
2. WHO. *Dengue Fact Sheet* (WHO, Geneva, 2018).
3. Guzman, M. G. & Harris, E. Dengue. *Lancet (London, England)* **385**, 453–465 (2015).
4. Tatem, A. J., Hay, S. I. & Rogers, D. J. Global traffic and disease vector dispersal. *Proc. Natl. Acad. Sci. USA* **103**, 6242–6247 (2006).
5. Wesolowski, A. *et al.* Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl. Acad. Sci.* **112**, 11887–11892 (2015).

6. Stanaway, J. D. *et al.* The global burden of dengue: An analysis from the Global Burden of Disease Study 2013. *Lancet. Infect. Dis* **16**, 712–723 (2016).
7. Halstead, S. B. Dengue vaccine development: A 75% solution?. *Lancet (London, England)* **380**, 1535–1536 (2012).
8. WHO. *Global Strategy for Dengue Prevention and Control 2012–2020* (World Health Organization, Geneva, 2012).
9. Dengue vaccine: WHO position paper, September 2018—Recommendations. *Vaccine* **37**, 4848–4849 (2018).
10. Lauer, S. A. *et al.* Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014. *Proc. Natl. Acad. Sci.* **115**, 201714457 (2018).
11. Reich, N. G. *et al.* Challenges in real-time prediction of infectious disease: A case study of Dengue in Thailand. *PLOS Negl. Trop. cal Dis.* **10**, e0004761 (2016).
12. Johansson, M. A., Reich, N. G., Hota, A., Brownstein, J. S. & Santillana, M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci. Rep.* **6**, 33707 (2016).
13. Yamana, T. K., Kandula, S. & Shaman, J. Superensemble forecasts of dengue outbreaks. *J. R. Soc. Interface* **13**, 20160410 (2016).
14. Promprou, S., Jaroensutasinee, M. & Jaroensutasinee, K. Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA models. *Dengue Bull.* **30**, 99–106 (2006).
15. Choudhury, Z., Banu, S. & Islam, A. Forecasting dengue incidence in Dhaka, Bangladesh: A time series analysis. *Dengue Bull.* **32**, 29–37 (2018).
16. Hu, W., Clements, A., Williams, G. & Tong, S. Dengue fever and El Niño/Southern Oscillation in Queensland, Australia: A time series predictive model. *Occup. Environ. Med.* **67**, 307 (2010).
17. Gharbi, M. *et al.* Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors. *BMC Infect. Dis.* **11**, 166 (2011).
18. Yang, S. *et al.* Advances in using Internet searches to track dengue. *PLoS Comput. Biol.* **13**, e1005607 (2017).
19. Martinez, E. Z., Silva, E. A. A. & Fabbro, A. L. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Rev. Soc. Bras. Med. Trop.* **44**, 436–440 (2011).
20. Hii, Y. L., Zhu, H., Ng, N., Ng, L. C. & Rocklöv, J. Forecast of dengue incidence using temperature and rainfall. *PLoS Negl. Trop. Dis.* **6**, e1908 (2012).
21. Eastin, M. D., Delmelle, E., Casas, I., Wexler, J. & Self, C. Intra- and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in Colombia. *Am. J. Trop. Med. Hygiene* **91**, 598–610 (2014).
22. Baquero, O., Santana, L. & Chiaravalloti-Neto, F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PLoS ONE* **13**, e0195065 (2018).
23. Buczak, A. L. *et al.* Ensemble method for dengue prediction. *PLoS ONE* **13**, e0189988 (2018).
24. Olliaro, P. *et al.* Improved tools and strategies for the prevention and control of arboviral diseases: A research-to-policy forum. *PLOS Negl. Trop. Dis.* **12**, e0005967 (2018).
25. Scarpino, S. V., Meyers, L. & Johansson, M. A. Design strategies for efficient arbovirus surveillance. *Emerg. Infect. Dis.* **23**, 642–644 (2017).
26. Chretien, J.-P., Rivers, C. M. & Johansson, M. A. Make data sharing routine to prepare for public health emergencies. *PLoS Med.* **13**, e1002109 (2016).
27. Stolerman, L. M., Coombs, D. & Boatto, S. SIR-network model and its application to dengue fever. *SIAM J. Appl. Math.* **75**, 2581–2609 (2015).
28. Arino, J. & van den Driessche, P. A multi-city epidemic model. *Math. Popul. Stud.* **10**, 175–193 (2003).
29. Liu, K. *et al.* Spatiotemporal patterns and determinants of dengue at county level in China from 2005–2017. *Int. J. Infect. Dis.* **77**, 96–104 (2018).
30. Lloyd, A. L. & Jansen, V. Spatiotemporal dynamics of epidemics: Synchrony in metapopulation models. *Math. Biosci.* **188**, 1–16 (2004).
31. Lourenço, J. & Recker, M. Natural, persistent oscillations in a spatial multi-strain disease system with application to dengue. *PLoS Comput. Biol.* **9**, e1003308 (2013).
32. Lee, S. & Castillo-Chavez, C. The role of residence times in two-patch dengue transmission dynamics and optimal strategies. *J. Theor. Biol.* **374**, 152–164 (2015).
33. Luz, P. M., Mendes, B. V., Codeço, C. T., Struchiner, C. J. & Galvani, A. P. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *Am. J. Trop. Med. Hygiene* **79**, 933–939 (2008).
34. Stolerman, L., Maia, P. & Kutz, J. N. Data-driven forecast of dengue outbreaks in Brazil: A critical assessment of climate conditions for different capitals. arXiv:1701.00166 (2016).
35. Johansson, M. A. *et al.* An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci.* **116**, 24268–24274 (2019).
36. Ray, E. L., Sakrejda, K., Lauer, S. A., Johansson, M. A. & Reich, N. G. Infectious disease prediction with kernel conditional density estimation. *Stat. Med.* **36**, 4908–4929 (2017).
37. Nunes, M. R. *et al.* Air travel is associated with intracontinental spread of dengue virus serotypes 1–3 in Brazil. *PLoS Negl. Tropical Dis.* **8**, e2769 (2014).
38. Lourenço, J. & Recker, M. The 2012 Madeira dengue outbreak: Epidemiological determinants and future epidemic potential. *PLoS Negl. Tropical Dis.* **8**, e3083 (2014).
39. Stoddard, S. T. *et al.* House-to-house human movement drives dengue virus transmission. *Proc. Natl. Acad. Sci.* **110**, 994–999 (2013).
40. Zhu, G., Liu, J., Tan, Q. & Shi, B. Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China. *PLoS Negl. Tropical Dis.* **10**, e0004633 (2016).
41. Wesolowski, A., O'Meara, W., Eagle, N., Tatem, A. J. & Buckee, C. O. Evaluating spatial interaction models for regional mobility in Sub-Saharan Africa. *PLoS Comput. Biol.* **11**, e1004267 (2015).
42. Limkittikul, K., Brett, J. & L'Azou, M. Epidemiological trends of dengue disease in Thailand (2000–2011): A systematic literature review. *PLoS Negl. Tropical Dis.* **8**, e3241 (2014).
43. Salje, H. *et al.* Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. *Proc. Natl. Acad. Sci.* **109**, 9535–9538 (2012).
44. Cummings, D. A. *et al.* Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature* **427**, 344–347 (2004).
45. Salje, H. *et al.* Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science* **355**, 1302–1306 (2017).
46. van Panhuis, W. G. *et al.* Region-wide synchrony and traveling waves of dengue across eight countries in Southeast Asia. *Proc. Natl. Acad. Sci.* **112**, 13069–13074 (2015).
47. Kraemer, M. U. G. *et al.* Mapping global variation in human mobility. *Nat. Hum. Behav.* **4**, 800–810 (2020).
48. Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P. & Tatem, A. J. High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **8**, e55882 (2013).
49. NESDB. *Gross Regional and Provincial Product Chain Measures 2015* (National Economic and Social Development Board of Thailand, Bangkok, 2017).
50. Luxen, D. & Vetter, C. Real-time routing with OpenStreetMap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2011) https://doi.org/10.1145/2093973.2094062.

51. Tinbergen, J. *Shaping the World Economy: Suggestions for an International Economic Policy* (Twentieth Century Fund, New York, 1962).
52. Lewer, J. J. & den Berg, H. A gravity model of immigration. *Econ. Lett.* **99**, 164–167 (2008).
53. Lu, F. S., Hattab, M. W., Clemente, C., Biggerstaff, M. & Santillana, M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nat. Commun.* **10**, 147 (2019).
54. Reich, N. G. *et al.* Case study in evaluating time series prediction models using the relative mean absolute error. *Am. Stat.* **70**, 285–292 (2016).
55. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2016).
56. Team R. C. R. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2018).

Acknowledgements

RJM and NE were supported by Asian Development Bank TA-8656. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders. MS was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM130668. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. COB and MS thank the Harvard Data Science Initiative for their support in partially funding this collaborative work.

Author contributions

C.O.B. conceptualized the study. M.V.K., C.O.B., and M.S. designed the methodology. K.E.-M., N.E., D.A., P.P., and R.J.M. curated the data. M.V.K. conducted all analyses. M.V.K., M.S., J.T.C., N.K., and C.O.B. interpreted the results. M.V.K. prepared the original draft. M.V.K., M.S., J.T.C., J.P.O., N.K., K.E.-M., N.E., D.A., P.P., R.J.M., and C.O.B. provided critical feedback. C.O.B. and R.J.M. supervised this work. M.V.K., M.S., J.T.C., J.P.O., N.K., K.E.-M., N.E., D.A., P.P., R.J.M., and C.O.B. reviewed and approved the submitted manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79438-0.

Correspondence and requests for materials should be addressed to C.O.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.