Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 81 (2018) 16–30

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison

Yirong Chen[a], Collins Wenhan Chu[b], Mark I.C. Chen[a,c], Alex R. Cook[a,*]

[a] Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Tahir Foundation Building, 12 Science Drive 2, 117549, Singapore

[b] Genome Institute of Singapore, 60 Biopolis Street, Genome, 138672, Singapore

[c] Department of Clinical Epidemiology, Communicable Disease Centre, Tan Tock Seng Hospital, Singapore, Moulmein Road, 308433, Singapore

ARTICLE INFO

Keywords:
LASSO
Endemic infectious disease
Real time forecast

ABSTRACT

Introduction: Accurate and timely prediction for endemic infectious diseases is vital for public health agencies to plan and carry out any control methods at an early stage of disease outbreaks. Climatic variables has been identified as important predictors in models for infectious disease forecasts. Various approaches have been proposed in the literature to produce accurate and timely predictions and potentially improve public health response.

Methods: We assessed how the machine learning LASSO method may be useful in providing useful forecasts for different pathogens in countries with different climates. Separate LASSO models were constructed for different disease/country/forecast window with different model complexity by including different sets of predictors to assess the importance of different predictors under various conditions.

Results: There was a more apparent cyclicity for both climatic variables and incidence in regions further away from the equator. For most diseases, predictions made beyond 4 weeks ahead were increasingly discrepant from the actual scenario. Prediction models were more accurate in capturing the outbreak but less sensitive to predict the outbreak size. In different situations, climatic variables have different levels of importance in prediction accuracy.

Conclusions: For LASSO models used for prediction, including different sets of predictors has varying effect in different situations. Short term predictions generally perform better than longer term predictions, suggesting public health agencies may need the capacity to respond at short-notice to early warnings.

1. Introduction

Outbreaks such as those caused by the Severe Acute Respiratory Syndrome Coronavirus (SARS CoV), the influenza A(H1N1)pdm09 pandemic of 2009, and more recently the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), Ebola virus and Zika virus have demonstrated the high potential risk of emerging and re-emerging infectious diseases to spread within and between countries [1–5]. These in turn cause increasing challenges for public health systems, including the increasing burden of infectious disease, and the need to build a surveillance and response system that is able to identify newly emerging disease rapidly, both regionally and internationally which calls for international collaboration, and the need for drug and vaccine research and production [6–8]. While the response to endemic diseases may be less urgent, the burden caused by pathogens such as influenza or malaria is high [9–11], and due to their endemicity, many countries have long standing surveillance systems to track outbreaks and guide response, from vector control to hospital bed utilization [12–15]. Early warning systems aiming to predict epidemics as soon as possible can allow control methods to be carried out rapidly and increase their chance of success [16,17]. To do so, decision makers need to be able to make accurate forecasts of incidence and to automate this forecasting process based on routinely collected notification data [18]. If accurate forecasts were available in both the near and far future, effective policies could then be targeted to the expected future needs. Existing approaches to real-time forecasting include generalized linear regression, seasonal autoregressive integrated moving average (SARIMA) model or a simpler ARIMA form of it, phenomenological models like the logistic growth model and Richards model, and mechanistic models like the SIR models [19–24]. Often such approaches involve the challenge of integrating environmental factors including temperature, humidity and rainfall, which may influence pathogen transmission directly or affect

* Corresponding author.
E-mail address: alex.richard.cook@gmail.com (A.R. Cook).

https://doi.org/10.1016/j.jbi.2018.02.014
Received 11 September 2017; Received in revised form 19 January 2018; Accepted 24 February 2018
Available online 27 February 2018
1532-0464/

the vector activities (for vector borne diseases), especially in temperate regions [25–29]. For instance, influenza virus is more transmissible in low temperature and low humidity conditions [30,31], while the primary vector of dengue, the yellow fever mosquito *Aedes aegypti*, favors higher temperature [32,33]. The availability of real time data-streams on seasonal variation and climatic variability therefore holds the potential to lead to more accurate prediction algorithms, potentially improving public health response.

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a machine learning method that can find patterns within large datasets while avoiding the problem of over-fitting [34]. Estimation and variable selection are simultaneously carried out using the LASSO method, and as such it is commonly used in studies in fields with large numbers of explanatory variables to reduce the variable space. This algorithm trades off model accuracy with model parsimony by introducing a penalty term into the objective function (which in standard linear regression is the sum of squares of residuals). The penalty term can, for linear regression models, be made equivalent to a constraint on the sum of the absolute parameter coefficients. This constraint imposed by LASSO regression has the effect of shrinking some estimated coefficients towards zero, which may help reduce biases caused by separation in some forms of regression [35], while simultaneously producing some parameter estimates that are exactly 0, so that the covariate associated with this coefficient is not associated with the outcome variable in that model. The optimal balance between model accuracy and complexity is typically obtained through cross-validation: repeatedly partitioning the data into training and validation sets, varying the degree of penalty, optimizing the regression parameters for each penalty value, then selecting the penalty that minimizes out of sample predictive accuracy. Computationally efficient methods to explore the penalty and parameter space exist [36], making it feasible to use LASSO as part of a 'real-time' forecasting pipeline for routinely collected health data such as infectious disease notifications. This computational speed allows the forecasting to adapt to changes in the underlying disease dynamics by permitting refitting of the model each time new data are reported, which may be important for diseases in which the severity changes between outbreaks, such as influenza [37]. Forecasts at different time horizons can be obtained through splicing together separate LASSO models, each trained on the data available at the time of the forecast, but tailored to predict at different windows into the future.

The LASSO method has previously been used in dengue outbreak prediction in Singapore, where it is now routinely used to guide vector control policy [38]. The objective of this paper is to apply the LASSO method to infectious disease forecasting and assess more generally in which situations LASSO models will provide useful forecasts. Unlike conventional use of the LASSO method to variable selection, the primary interest of our application of the LASSO-based method on infectious disease data is to make forecast of incidences in the future, rather than to identify the potential risk factors to explain the epidemics of these infectious diseases.

In particular, we assess for diseases with different transmission modalities, in different climatic zones, how accurate short to medium time forecasts can be, and what data streams are necessary for accurate forecasts. We apply the method to four countries from different latitudes—Japan, Taiwan, Thailand and Singapore—to cover temperate, sub-tropical and tropical settings.

2. Methods

2.1. Sources of data

Four representative countries with distinct climates were selected for analysis based on Köppen-Geiger climate classification [39] – Japan with humid continental and subtropical climate; Taiwan with humid subtropical and oceanic climate; Thailand with tropical wet and savannah climate and Singapore with tropical rainforest climate. Four representative infectious diseases were included in the study: two mosquito-borne infections (Dengue and Malaria) and two infections that spread from person to person (Hand Foot and Mouth Disease (HFMD) and Chickenpox). For all four pathogens, a relationship has previously been found between incidence and climatic variables [40–42] or for there to be a seasonality to incidence [43]. Not all four pathogens were considered for each country: some are not present in each country while others are not captured in routine infectious disease surveillance systems.

The notified numbers of chickenpox, HFMD cases in Japan were collected by the National Institute of Infectious Diseases (NIID) [44]. Both were reported as average cases per week per sentinel reporting, to accommodate varying reporting rates. We extracted weekly data from 2001 to 2012.

Monthly reported cases of chickenpox, dengue, and malaria in Thailand for the period 2003–2013 were obtained from the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand [45]. The number of incident cases were collected from government hospitals, public health offices and health centers by the National Disease Surveillance [46] and were reported online.

Ministry of Health, Singapore, actively monitors and publishes the incidence of dengue and HFMD in Singapore, both being notifiable diseases. Weekly number of incidences for the period 2003 and 2014 were obtained from the Weekly Infectious Diseases Bulletin [47].

Weekly number of dengue cases from 2003 to 2014 were extracted from Taiwan National Infectious Disease Statistics System [48]. Both indigenous and imported cases were included in the count.

Epidemiological week as per US Centers for Disease Control and Prevention was used in our analysis using the EpiWeek package in R [49].

Climatic data for Taiwan, Thailand and Singapore were obtained from the Weather Underground [50] which documented among other variables, historical temperature, humidity, sea level pressure, and visibility. Only temperature (daily highest, average and lowest) and relative humidity (daily highest, average and lowest) were used in our models due to insufficient historical data of other climatic variables. Climatic data for Japan were obtained from the Japan Meteorological Agency [51], which provides and archives various weather information. Weekly mean temperature, relative humidity and rainfall information were used in our model. For all locations, the weather data at the capital (Tokyo, Taipei, Bangkok and Singapore) was used to represent overall national weather.

2.2. Statistical analysis

Wavelet analyses were done to explore periodicity of all endemic diseases and climatic variables in four countries. The wavelet approach was based on a wavelet function which analyses locality in time and frequency [52]. Wavelet transformation ($W_t(s)$) as the convolution of the time series x_t with Morlet function $\psi_0(\eta)$ at scale s was conducted:

$$W_t(s) = \sum_{k=0}^{T-1} \hat{x}_k \hat{\psi}^*(s\omega_k) e^{i\omega_k n\delta_t}$$

where $\hat{x}_k$ is the discrete Fourier transform of x_t: $\hat{x}_k = \frac{1}{T}\sum_{t=0}^{T-1} x_t e^{-2\pi ikt/T}$, $k = 0,...,T-1$ the frequency index, and $\hat{\psi}(s\omega)$ the Fourier transform of Morlet function: $\hat{\psi}(s\omega) = \pi^{-\frac{1}{4}} H(\omega) e^{-(s\omega-\omega_0)^2/2}$, where ω_0 refers to the nondimensional frequency and is set to 6 to satisfy the admissibility condition [53].

The wavelet transformation $W_t(s)$ can be divided into amplitude, $|W_t(s)|$, and phase, $\tan^{-1}[\Im\{W_n(s)\}/\Re\{W_n(s)\}]$, where $\Re\{W_n(s)\}$ is the real part of the transform and $\Im\{W_n(s)\}$ the imaginary part. The wavelet power spectrum is defined as $|W_t(s)|^2$ [54].

Least Absolute Shrinkage and Selection Operator (LASSO) regression [34] is used for the prediction models. As incidence data were generally right skewed and with increased stochasticity during epidemics, we controlled for skewness and heteroscedasticity by log-transforming (after adding one to) the observed number of cases for all analyses, or mean number of cases per sentinel for Japan.

We build separate LASSO models for each disease, in each country, and for each forecast window based on similar sets of candidate predictors (adapting the approach used in Ref. [38]). Each candidate predictor appeared several times in each model, at different historical lags. To allow for contagion and typical epidemic duration, we used past incidence of up to a half year (26 weeks for Japan, Taiwan and Singapore; 6 months for Thailand) as autocovariates. We accommodated non-linearities in the relationship between past and future incidence by using the quadratic of past incidence (again, up to one half year in the past) as a potential covariate; the inclusion of additional polynomial terms did not improve the results and hence were omitted. We used climatic variables (minimum, maximum and average temperature, humidity and rainfall) of up to 4 weeks (one month for Thailand). We also used a month indicator that associated each epidemiological week with a single calendar month to represent potential seasonal variation beyond those governed by the climatic data. Wavelet transformations of the covariates and outcomes were not used, to ensure that only data available at the time of the forecast were used to make the forecast.

For each forecast window (from $k = 1$ time unit to 6, for Thailand (the data being monthly) or 26, for Japan, Taiwan and Singapore (weekly)), a separate LASSO model was developed which used data available at the time of the forecast only. The selection of the time window was dependent on the resolution of notification data for each disease. Each LASSO model is as follows:

$$\begin{aligned} y_{t+k} = {} & \alpha_{t_k} + \beta_{t_k} y_t + \beta_{t_k-1} y_{t-1} + \ldots + \beta_{t_k-25} y_{t-25} + \gamma_{t_k} y_t^2 + \gamma_{t_k-1} y_{t-1}^2 \\ & + \ldots + \gamma_{t_k-25} y_{t-25}^2 + \delta_{t_k} c_t + \ldots + \delta_{t_k-3} c_{t-3} + \theta_{k_1} \mathrm{Jan}_{t+k} + \theta_{k_2} \mathrm{Feb}_{t+k} \\ & + \ldots + \theta_{k_{12}} \mathrm{Dec}_{t+k} + \varepsilon_{t_k} \end{aligned}$$

where y_t is the log number of cases at time t, y_t^2 the square of the log number of cases at time t, c_t the climatic variable at time t, and Jan_t, Feb_t, …, Dec_t are indicator variables for the month at time t.

This was subject to the constraint

$$\sum_{i=0}^{25} |\beta_{t_k-i}| + \sum_{i=0}^{25} |\gamma_{t_k-i}| + \sum_{i=0}^{3} |\delta_{t_k-i}| + \sum_{i=1}^{12} |\theta_{k_i}| \leqslant p.$$

The p and penalty term λ have a one-to-one correspondence in LASSO, which can equivalently be formulated as a constraint to the total absolute size of the parameters, or as a penalty to the log-likelihood determined by their total absolute size. The choice of the optimal p in this study was based on ten-fold cross validation for that dataset (country, disease, forecast length combination). Ten-fold cross validation was carried out by partitioning the total data set into 10 approximately equal-sized subsamples. Each subsample was used in turn as the validation set while the rest were used as the training set, the performance was measured through the mean squared error (of the difference between the forecast and the data) on each validation set and averaged over the ten folds. The constraint (or equivalently, penalty term) that optimized the out of sample performance was then used to refit the model and make projections. In this way, the optimal complexity of the model—from the perspective of out of sample prediction—can be determined via the constraint. We developed both (a) 'real time' forecasts, in which both the parameters and the predictions were regenerated based on available data at that time, and (b) one set of 'retrospective' forecasts, which used the last of the real time forecasts, to summarize effects of covariates.

Table 1
Sets of predictors in different models.

Models	Incidence	Squared incidence	Climatic variables	Monthly effect
1	√	.	.	.
2	√	√	.	.
3	√	.	√	.
4	√	.	.	√
5	√	√	√	.
6	√	√	.	√
7	√	.	√	√
8	√	√	√	√

Following standard practice, covariates were z-scored prior to estimation and the coefficients rescaled afterwards.

As the models were built separately for each forecast window, the variables selected and their lags and parameter magnitude and sign may differ substantially.

2.3. Sensitivity analysis

Eight models with different combination of predictor types were used for each disease/country/forecast length to assess the importance of including each datastream in the predictive algorithm. Details of the components in each model are shown in Table 1.

2.4. Accuracy

To assess prediction accuracy of the models, we made out-of-sample forecasts and compared with data observed. For each time point and each prediction window, a 95% projection interval—which we define here as an interval derived from the best fitting model with random errors overlaid, but excluding parametric uncertainty—was derived and the percentage of such intervals successfully capturing the actual value was calculated.

To assess if the prediction models were able to detect early signs of elevated levels of transmission—which we defined to be time windows (weeks or months) in which the number of cases is more than the 75th percentile of what has been observed in the preceding year—we calculated the fraction of predictions in which the projection correctly classified the data as being above or below that threshold. This was calculated separately for each length of forecast window. An accuracy of 75% is expected for a poorly predictive model. One-sided binomial tests were done to all prediction windows and those that significantly identify the early signs at an accuracy of more than 75% were identified.

Prediction accuracy was measured by the mean absolute percentage error (MAPE), root mean squared error (RMSE), and R-squared, for each forecast window. Relative prediction error as compared to the simplest incidence only model at all forecast windows was calculated to assess model complexity and accuracy. All statistical analysis were performed using R Statistical Software [55].

3. Results

Climatic and incidence patterns of the diseases are presented in Fig. 1. Four representative diseases were shown in this figure (others are presented in Supplementary Figs. 1–4). At further distances from the equator, temperature and humidity show more obvious annual cycles, and there is concomitantly more cyclicity in disease incidence, as indicated in the power spectra illustrated in Fig. 2. Power spectra of average temperature and humidity are presented in Supplementary Figs. 5 and 6. Temperature in all countries has a period of one year but the cyclic effect is more marked in temperate and subtropical regions,

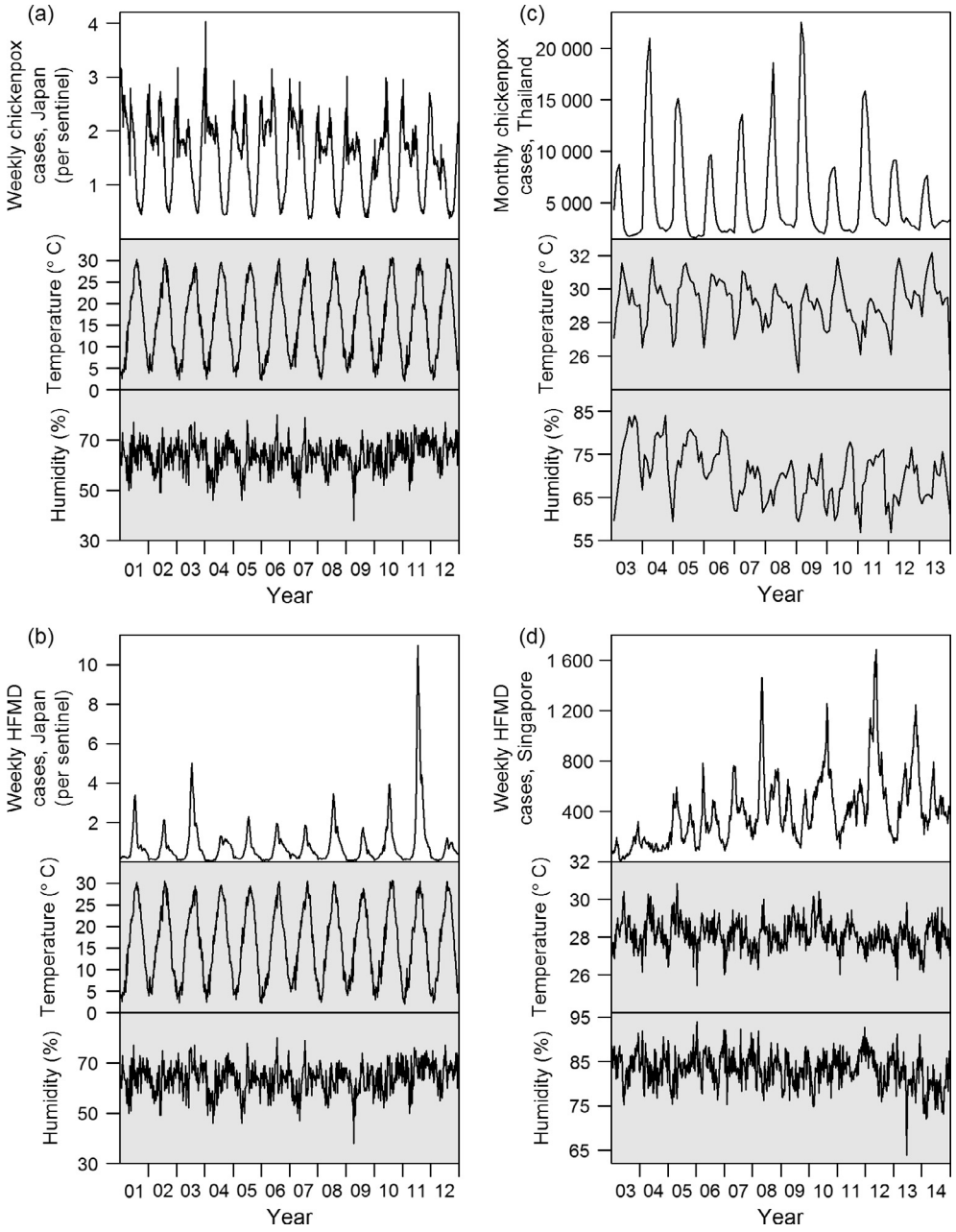
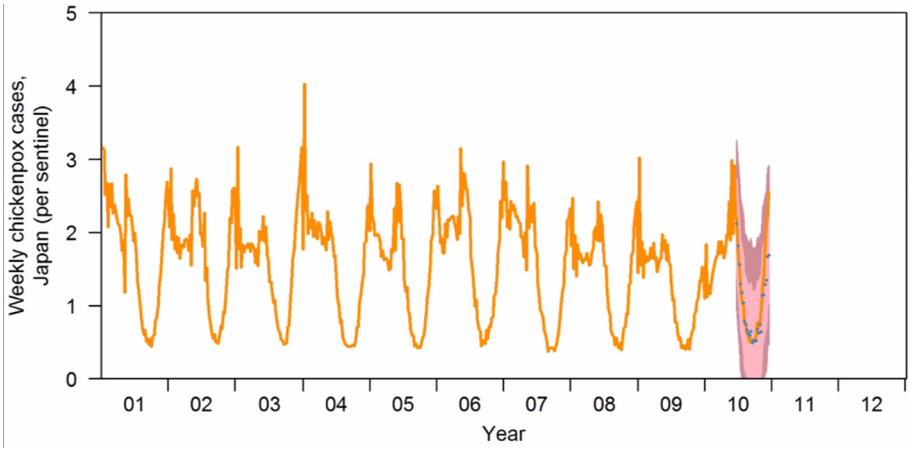


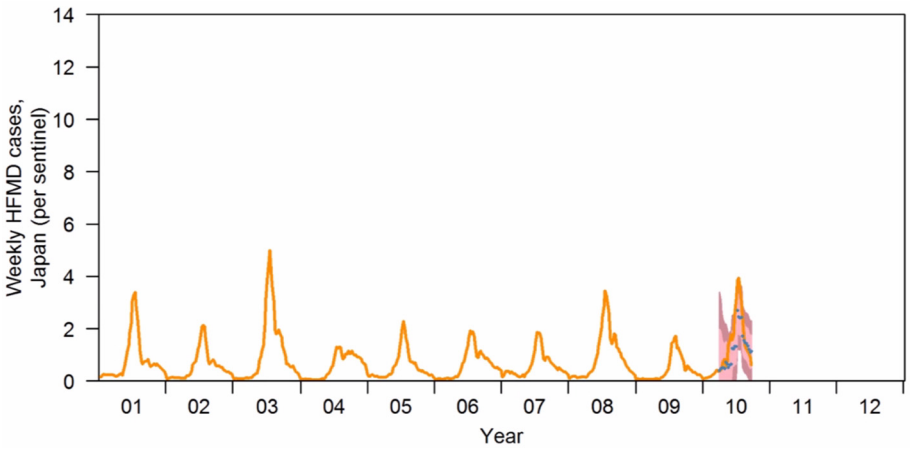
Fig. 1. Incidence and climatic (temperature and humidity) patterns of four representative diseases. (a) Chickenpox in Japan; (b) HFMD in Japan; (c) chickenpox in Thailand; (d) HFMD in Singapore. Year 01 corresponds to the year 2001 of the common era, et cetera.

and less so in equatorial Singapore. Average humidity displays one-year cycle in temperate and subtropical regions, but again not Singapore, although the effect is not as obvious as with average temperature. Although in temperate regions, there is an apparent synchronicity between incidence and climatic factors, in tropical regions like equatorial Singapore, there was little variation in either temperature or humidity, and similarly no observed cycle of disease incidence. In addition, there was no correlation observable between cases and climate, either in incidence itself or the change in incidence (reflecting transmissibility).

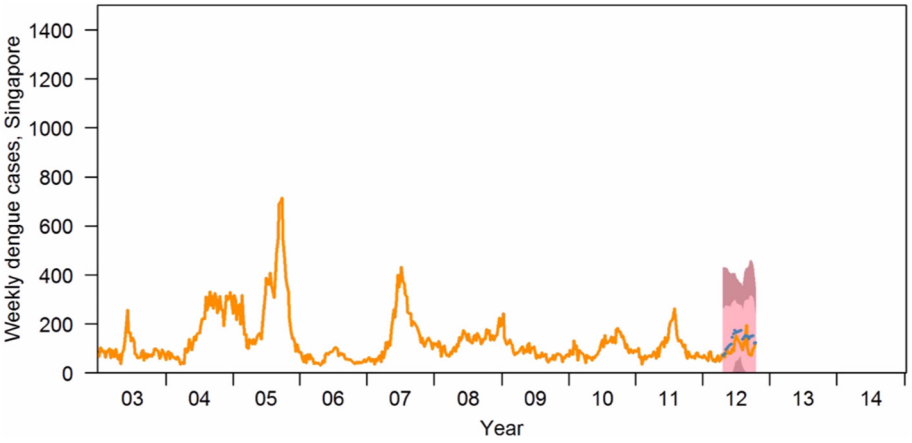
Fig. 3 shows the actual incidences as well as forecasts at four prediction points for the four diseases presented in Fig. 1 (predictions at other time points/diseases are presented in Supplementary Videos 1–8) using the model with all variable sets. At each time point, up to six month ahead predictions are shown in the figure, using data available at the time the prediction was made. The number of variables from each category that were selected by our LASSO models are presented in Supplementary Fig. 7 and the effect size of each climatic variable at different lags for all forecast windows are shown in Supplementary Fig. 8. Generally, the 95% projection interval of the forecast captured the actual scenario more than 80% of the time, indicating slight undercoverage. Coverage rates for the full models are tabulated in Table 2. MAPE for full models at all prediction windows are shown in Table 3. RMSE and R-squared values for full models are included in Supplementary Tables 1 and 2. For most diseases, predictions made beyond 4 weeks ahead were increasingly discrepant from the actual scenario.



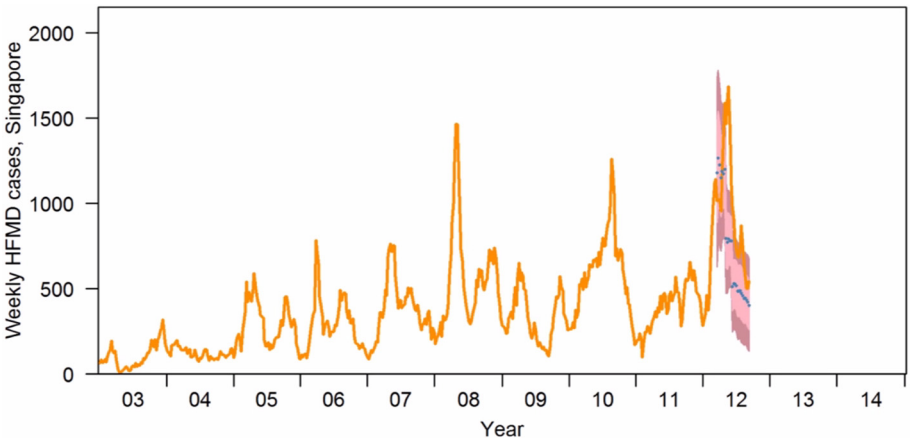
Video 1.



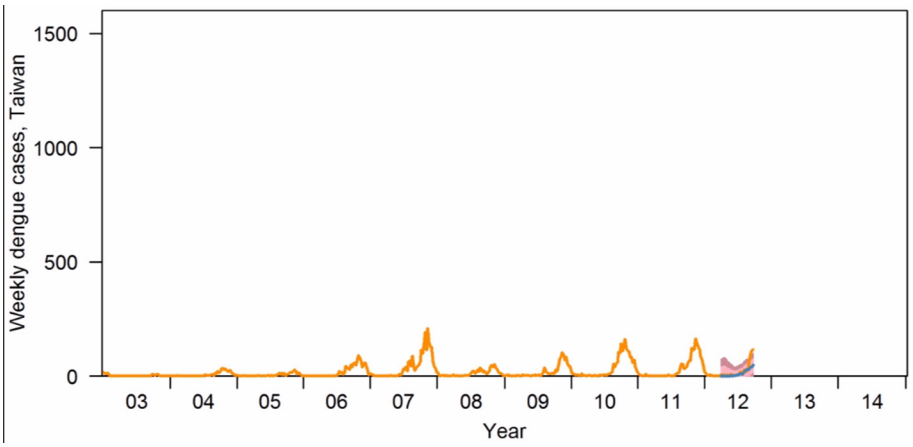
Video 2.



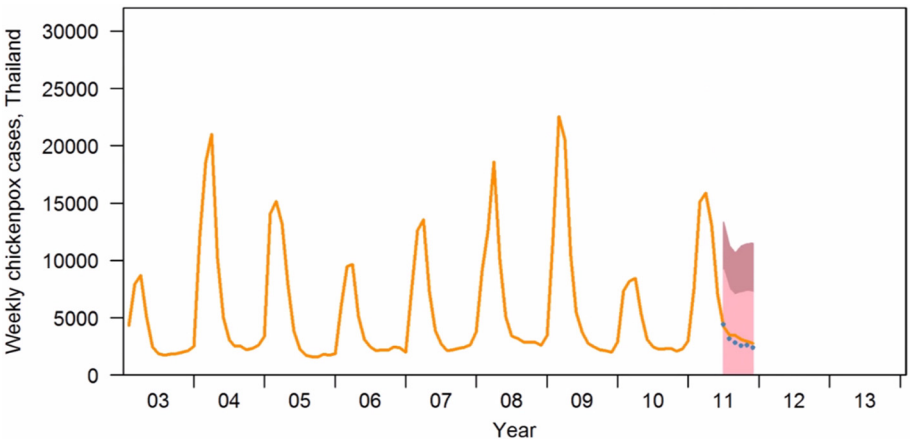
Video 3.



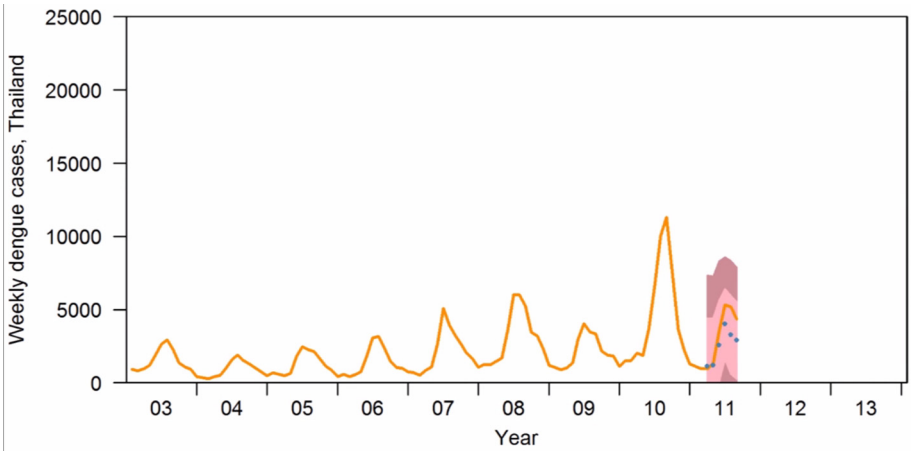
Video 4.



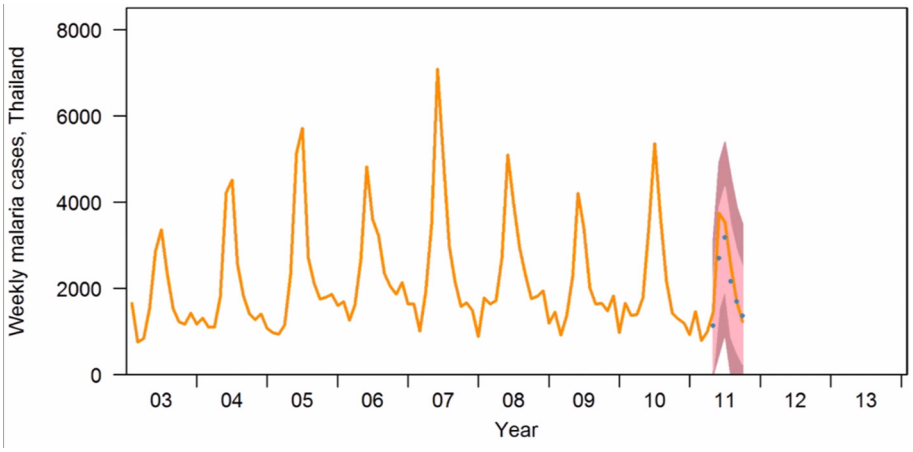
Video 5.



Video 6.



Video 7.



Video 8.

Predicted cases against actual cases at 1 week, 2 weeks, 4 weeks and 8 weeks for all prediction period for all diseases were plotted in Fig. 4. For all diseases, as the prediction window gets longer, the magnitude of the error grows, especially for the largest sizes of outbreaks where there is likely to be more stochastic variability. Cumulative density functions of the observed time series of all diseases and predicted time series at various forecast windows are shown in Fig. 5: those derived from predictions are generally close to that from the observed time series. Short term predictions generally performed better than longer term predictions. Table 4 shows the classification accuracy of our prediction models being able to correctly classify one week/month to be an epidemic week/month. A percentage above 75% suggested that the prediction model performed better than a random guess. Models that passed the binomial test of having accuracy significantly higher than 75% were marked with the asterisk. Predictions for HFMD epidemic were generally accurate even at long prediction windows; for most other diseases, prediction within 5 weeks were above 75%. However, for chickenpox in Japan and dengue in Singapore, only predictions within 3 and 4 weeks were significantly above 75% from the Binomial test. This suggests that the approach can predict the timing of periods of elevated transmission at about one month in advance but that the absolute magnitude is hard to predict accurately, given the larger errors at high levels exhibited in Fig. 4.

The trade-off between model complexity and efficiency was assessed by constructing models with different groups of variables and comparing the forecasts based on these models (Fig. 6, an enlarged version is shown in Supplementary Figs. 9 and 10). The prediction errors arising from each model were then compared with the simplest incidence only model. A ratio substantially less than 1 shows prediction accuracy increased by including the set(s) of variables. For some diseases like chickenpox in Thailand, dropping some variables affected prediction accuracy by a large extent, indicating the importance of including monthly effect to achieve greater accuracy in predictions. For other diseases like dengue in Taiwan, dropping the climatic variables had only minimal effect on the accuracy of the predictions, which suggests that the features in those sets are not necessary for routine predictions. Panels in the second and fourth columns in Fig. 6 showed prediction error for all diseases for different prediction windows. Prediction error was quantified by the mean absolute percentage error. In line with our previous observations, as the prediction window gets longer, prediction error increases. More complicated models show less prediction error but the improvements varied by diseases and countries.

4. Discussion

Applying the LASSO method to data on recent incidence and climate leads to around 20% or lower prediction error for short term predictions, even for an aseasonal, tropical country like Singapore in which the presence of randomness in outbreak occurrence may be strongly felt [56]. Most of the variation in incidence was explained by the autoregressive terms as there was no or minimal improvement in forecast accuracy by including more variables other than autoregressive terms in the prediction model shown in Fig. 6. Long term predictions were not accurate, even when we include multiple climatic variables as candidate features. We believe this reflects the difficulty in using weather data relating to current conditions to predict disease levels in several months—in contrast, most other analyses use climate data from near the time point being forecast [28,57], which is typically not available in

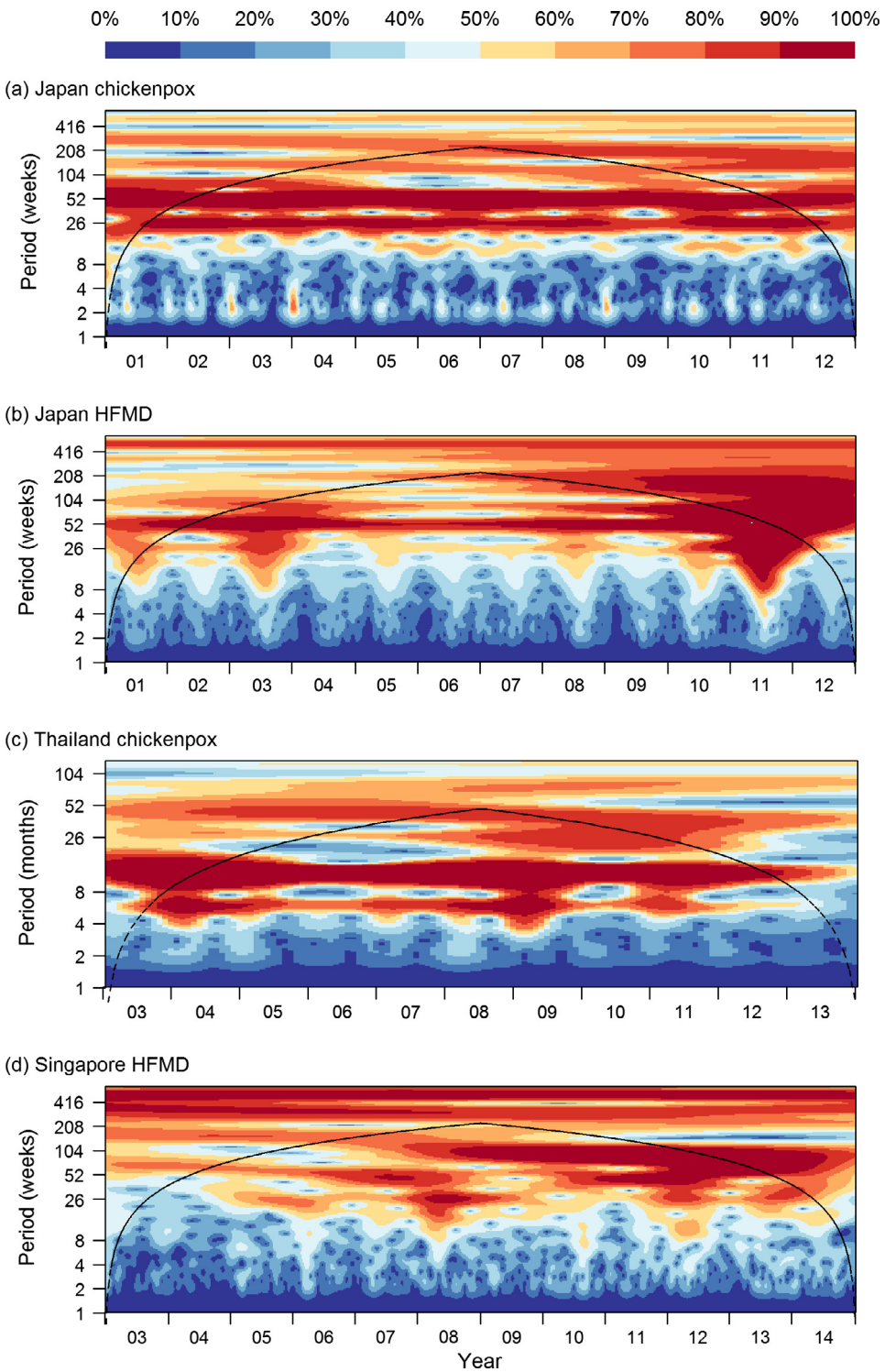


Fig. 2. Wavelet power spectrum for (a) chickenpox in Japan; (b) HFMD in Japan; (c) chickenpox in Thailand; (d) HFMD in Singapore. For each panel, the power spectrum values are categorised by decile prior to plotting.

real time. Hence the utility of predictive analytics for infectious disease policy making is mostly for immediate response over the next few weeks rather than months.

This LASSO-based method we proposed for forecast provides an insight into the potential utilization of LASSO regression outside its usual application in the area of variable selection, for real time forecasting. This approach can readily be automated to run on routinely collected case data, but the accuracy of the prediction is very much dependent on the timeliness at which such data become available. This calls for a comprehensive regional and country level infectious diseases surveillance system, whether using clinician/laboratory driven case notification, or mining of electronic medical records. Weather data, in

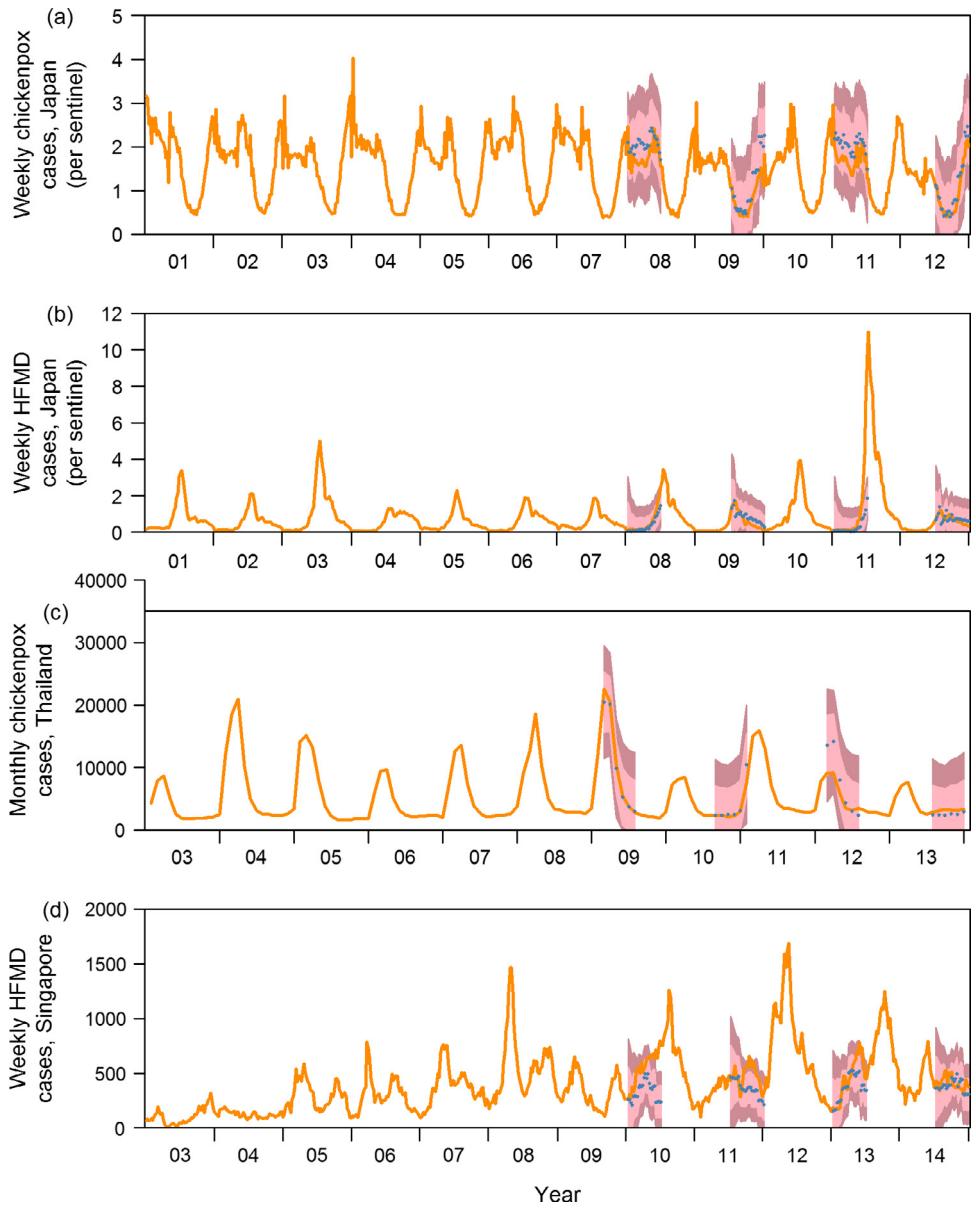


Fig. 3. Actual incidence (orange line) and forecasts (blue dots, dark red 95% projection interval, pink 70% projection interval) at four time points for each of the representative diseases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

contrast, are widely posted to sources like the Weather Underground which can be mined in real-time. An example of this in practice can be found in Singapore, where the National Environment Agency has a weekly dengue forecast, as described elsewhere [38]. The LASSO model building process involves identifying the model complexity through optimal out of sample predictive performance, and as such it is naturally suited to making forecasts. By fitting separate models at each forecast window, variables whose predictive performance is expressed differently at different lags can be efficiently exploited: this might be particularly beneficial for vector borne diseases in which the effect of weather on transmission is mediated through one or more generations of vectors.

In large countries, and in particular those that span multiple latitudes, we anticipate that the accuracy of the predictions will improve if regional information rather than aggregated country-level data streams are used. The aggregation averages the whole country and thus attenuates regional differences in both climate and incidence. Future work could assess the impact of disaggregation for countries like Japan for which prefectural level data are available.

By using past disease incidence and seasonal terms, LASSO has many of the characteristics of traditional time series models (e.g. the SAR in SARIMA) which have frequently been used for prediction of infectious disease time series [20,58–61]. What is less clear is how to incorporate structural information about outbreak data, such as the typical Gaussian incidence curve that is seen in both empirical data [62] and theoretical transmission dynamic models [63], or the known impact of changing herd immunity levels [64]. This may also help predict atypical outbreaks, such as the 2009 influenza pandemic which occurred in the Northern Hemisphere's summer [65], and for which extrapolating approaches like LASSO may break down.

5. Limitations

There are several limitations to this study. Regression models,

Table 2
Fraction of 95% projection intervals that cover the eventual value for full models for all diseases at all prediction windows.

	Japan		Singapore		Taiwan	Thailand		
Prediction Window	Chickenpox	HFMD	Dengue	HFMD	Dengue	Chickenpox	Dengue	Malaria
1	100%	100%	100%	100%	99%			
2	100%	98%	100%	100%	98%			
3	100%	96%	99%	98%	97%			
4	100%	96%	98%	96%	96%	100%	100%	100%
5	100%	95%	98%	94%	94%			
6	100%	94%	97%	87%	92%			
7	100%	92%	95%	83%	91%			
8	100%	91%	93%	77%	89%	100%	98%	96%
9	100%	91%	91%	74%	86%			
10	100%	90%	90%	73%	85%			
11	100%	90%	88%	71%	82%			
12	100%	89%	86%	69%	80%	100%	98%	96%
13	100%	89%	83%	64%	80%			
14	100%	89%	84%	63%	81%			
15	100%	90%	83%	63%	81%			
16	100%	90%	83%	63%	81%	100%	94%	100%
17	100%	91%	89%	65%	80%			
18	100%	90%	91%	64%	80%			
19	100%	90%	92%	64%	80%			
20	99%	90%	91%	64%	80%	100%	93%	98%
21	99%	90%	90%	65%	81%			
22	99%	90%	91%	63%	82%			
23	99%	90%	90%	63%	82%			
24	99%	90%	89%	63%	81%	100%	94%	98%
25	100%	90%	86%	63%	82%			
26	100%	90%	82%	63%	81%			

including those built using LASSO, may be effective in providing accurate predictions, but differ from traditional epidemic models such as susceptible-infected-removed compartmental models, network models [66] or individual based simulation [67], which seek to describe the transmission dynamics of contagious diseases, and can therefore be used to predict the effect of interventions or novel scenarios. A further limitation is that, for simplicity, only weather information from the capitals were used in the study, but for countries like Japan, Thailand and Taiwan, there are great variation between climatic conditions in the North and the South. Prediction accuracy might improve if

Table 3
Mean Absolute Percentage Error for full models for all diseases at all prediction windows.

	Japan		Singapore		Taiwan	Thailand		
Prediction Window	Chickenpox	HFMD	Dengue	HFMD	Dengue	Chickenpox	Dengue	Malaria
1	0.10	0.18	0.17	0.12	0.47			
2	0.10	0.31	0.20	0.18	0.57			
3	0.13	0.40	0.23	0.23	0.64			
4	0.14	0.46	0.26	0.26	0.66	0.09	0.15	0.20
5	0.16	0.51	0.29	0.29	0.67			
6	0.18	0.58	0.31	0.31	0.70			
7	0.19	0.65	0.33	0.33	0.70			
8	0.20	0.69	0.35	0.35	0.74	0.17	0.24	0.29
9	0.22	0.69	0.36	0.38	0.75			
10	0.22	0.68	0.36	0.41	0.77			
11	0.23	0.75	0.37	0.43	0.80			
12	0.23	0.83	0.39	0.45	0.80	0.24	0.25	0.31
13	0.23	0.89	0.38	0.46	0.82			
14	0.25	0.89	0.39	0.47	0.80			
15	0.25	0.93	0.39	0.48	0.79			
16	0.25	0.95	0.40	0.49	0.80	0.28	0.30	0.31
17	0.25	0.92	0.42	0.51	0.80			
18	0.25	0.91	0.42	0.51	0.79			
19	0.24	0.90	0.42	0.51	0.82			
20	0.24	0.92	0.42	0.53	0.82	0.28	0.32	0.30
21	0.24	0.92	0.43	0.53	0.79			
22	0.24	0.85	0.45	0.52	0.79			
23	0.23	0.69	0.46	0.51	0.78			
24	0.23	0.66	0.46	0.50	0.79	0.31	0.33	0.32
25	0.21	0.64	0.47	0.51	0.81			
26	0.22	0.60	0.46	0.51	0.81			

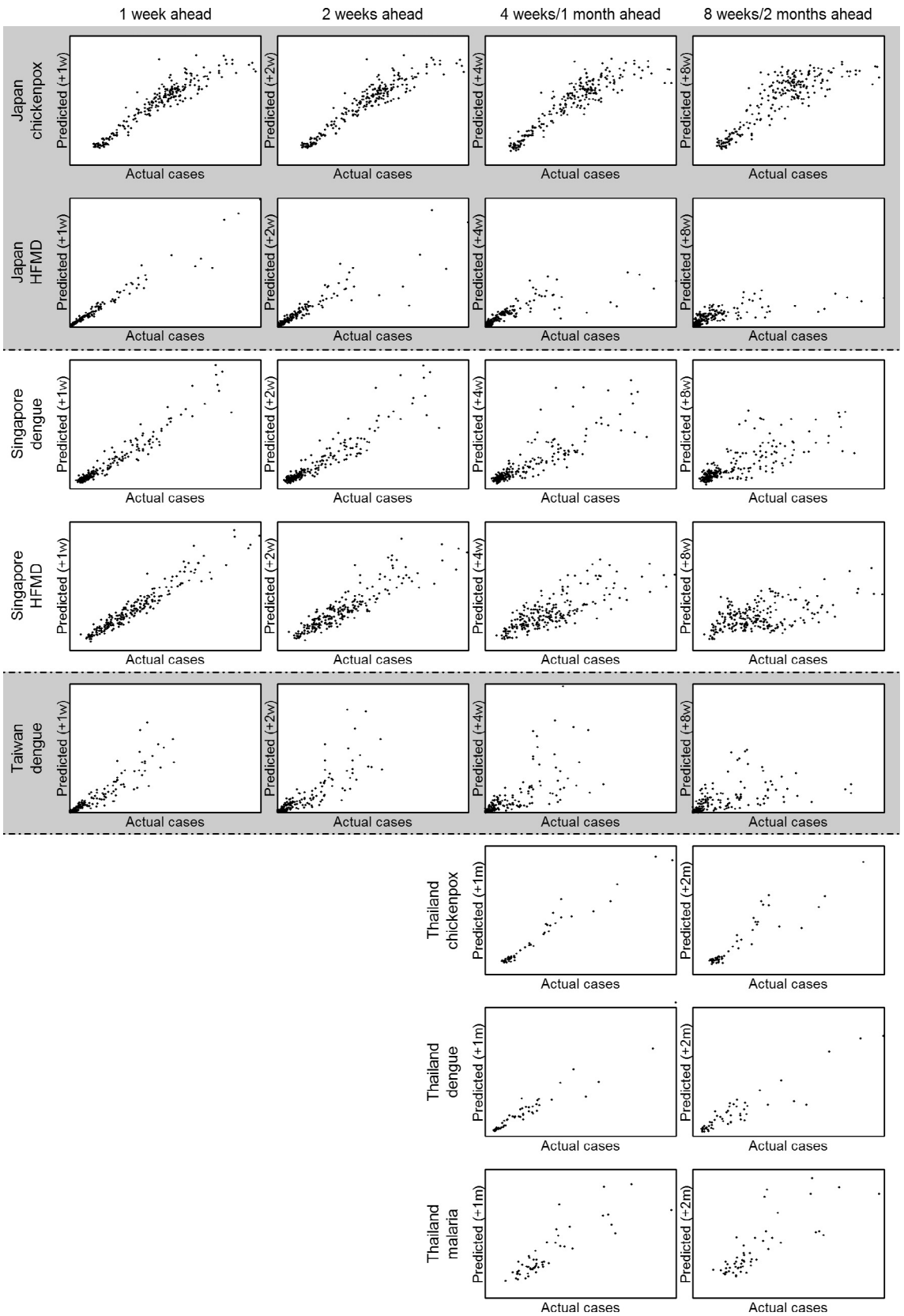


Fig. 4. Predicted cases against actual incidence at 1 week, 2 weeks, 4 weeks (1 month), and 8 weeks (2 months) for all prediction period for all diseases.

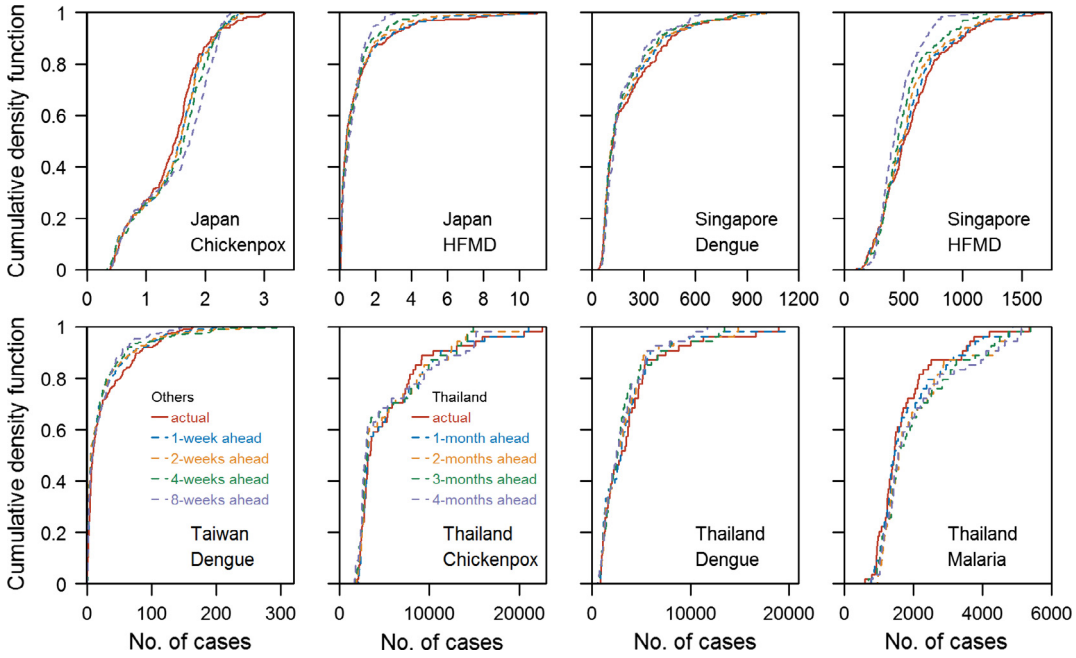


Fig. 5. Cumulative density functions (CDF) for observed time series for all diseases and the CDF for predicted time series at various forecast windows.

Table 4
Percentage of our prediction model correctly classify a predicted future week/month being an epidemic week/month. An epidemic week/month is defined as when the number of cases is more than the 75th percentile of the level in the past year. Percentage above 75% is shown bold. Models that had significantly higher accuracy than 75% are marked with an asterisk.

	Japan		Singapore		Taiwan	Thailand		
Prediction Window	Chickenpox	HFMD	Dengue	HFMD	Dengue	Chickenpox	Dengue	Malaria
1	**89%***	**97%***	**90%***	**99%***	**93%***			
2	**87%***	**97%***	**88%***	**96%***	**93%***			
3	**86%***	**97%***	**83%***	**95%***	**91%***			
4	**78%**	**95%***	**81%***	**94%***	**89%***	**98%***	**98%***	**100%***
5	**76%**	**93%***	**79%**	**91%***	**89%***			
6	75%	**91%***	**80%**	**90%***	**86%***			
7	72%	**89%***	75%	**87%***	**83%***			
8	71%	**88%***	73%	**85%***	**80%***	**94%***	**94%***	**100%***
9	68%	**88%***	69%	**81%***	**80%***			
10	69%	**88%***	71%	**80%***	**77%**			
11	70%	**88%***	70%	**79%**	75%			
12	69%	**88%***	71%	**80%***	75%	**93%***	**96%***	**100%***
13	71%	**87%***	70%	**81%***	74%			
14	70%	**87%***	68%	**81%***	**76%**			
15	69%	**87%***	69%	**81%***	**77%**			
16	71%	**86%***	71%	**80%***	**76%**	**91%***	**93%***	**100%***
17	73%	**86%***	70%	**80%***	**77%**			
18	72%	**88%***	73%	**80%***	**77%**			
19	73%	**89%***	71%	**80%***	**76%**			
20	71%	**90%***	71%	**80%***	**77%**	**89%***	**93%***	**100%***
21	71%	**90%***	73%	**79%**	**78%**			
22	70%	**89%***	73%	**79%**	**77%**			
23	68%	**89%***	72%	**80%***	**77%**			
24	70%	**89%***	71%	**80%***	**77%**	**87%***	**91%***	**100%***
25	71%	**89%***	71%	**78%**	**77%**			
26	73%	**89%***	71%	**81%***	**76%**			

incidence and weather information can be collected at a finer resolution.

We suspect that even the accuracy of short term forecasts may be reduced should new epidemiological conditions replace those that the model was trained on. Examples of this would be the emergence of a new strain or variant, such as an influenza pandemic [68], more virulent strain of the virus [69], or a virus moving into a new population [70]. It would also break down in the presence of changing control efforts such as school closure [71] or novel vector control [72]. In such situations, a mechanistic modelling approach [73] may be better able to predict the epidemic dynamics until sufficient training data are available.

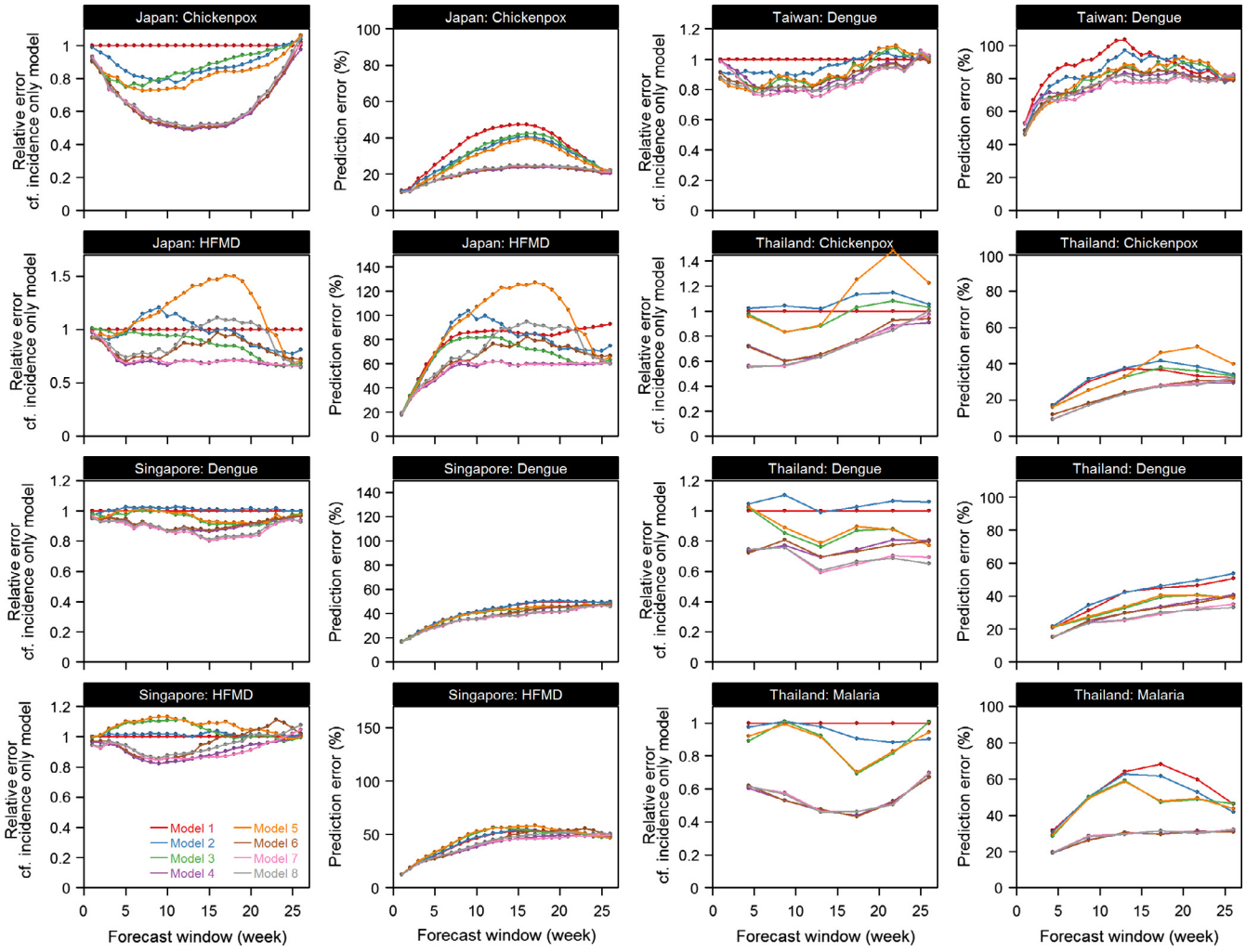


Fig. 6. Relative prediction error for models of different complexity as compared to the simplest incidence only model and absolute prediction error measured by MAPE for all models.

6. Conclusion

Regression models using LASSO were built to forecast incidence for four endemic diseases in four countries. For some diseases including one set of variables may improve predictive accuracy substantially while for other diseases, the simpler models give similar results as more complex models. For all diseases, short-term prediction were generally much better than longer term predictions, which suggests that public health agencies may need the capacity to respond at short-notice to early warnings of possible infectious disease outbreaks should models based on this approach be implemented routinely.

Funding

The work was partially supported by the Singapore Population Health Improvement Centre (SPHERIC), National University Health System.

Declarations of interest

None.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2018.02.014.

References

[1] D.L. Heymann, G. Rodier, Global surveillance, national surveillance, and SARS, Emerg. Infect. Dis. 10 (2) (2004) 173–175.
[2] Clinical aspects of pandemic 2009 influenza A (H1N1) virus infection, N. Engl. J. Med. 362(18) (2010) 1708–1719.
[3] A. Assiri, A. McGeer, T.M. Perl, C.S. Price, A.A. Al Rabeeah, D.A.T. Cummings, et al., Hospital outbreak of Middle East respiratory syndrome coronavirus, N. Engl. J. Med. 369 (5) (2013) 407–416.
[4] G. Rainisch, M. Shankar, M. Wellman, T. Merlin, M.I. Meltzer, Regional spread of Ebola virus, West Africa, 2014, Emerg. Infect. Dis. 21 (3) (2015) 444–447.
[5] D. Baud, D.J. Gubler, B. Schaub, M.C. Lanteri, D. Musso, An update on Zika virus infection, Lancet Lond. Engl. 390 (2017) 2099–2109, http://dx.doi.org/10.1016/S0140-6736(17)31450-2.
[6] D.M. Morens, G.K. Folkers, A.S. Fauci, The challenge of emerging and re-emerging infectious diseases, Nature 430 (6996) (2004) 242–249.
[7] L. Wang, Y. Wang, S. Jin, Z. Wu, D.P. Chin, J.P. Koplan, et al., Emergence and control of infectious diseases in China, The Lancet 372 (9649) (2008) 1598–1605.
[8] A.S. Fauci, Emerging and reemerging infectious diseases: the perpetual challenge, Acad. Med. J. Assoc. Am. Med. Coll. 80 (12) (2005) 1079–1085.
[9] N.-A.M. Molinari, I.R. Ortega-Sanchez, M.L. Messonnier, W.W. Thompson, P.M. Wortley, E. Weintraub, et al., The annual impact of seasonal influenza in the US: measuring disease burden and costs, Vaccine 25 (27) (2007) 5086–5096.
[10] S.K. Peasah, E. Azziz-Baumgartner, J. Breese, M.I. Meltzer, M.-A. Widdowson, Influenza cost and cost-effectiveness studies globally – a review, Vaccine 31 (46) (2013) 5339–5348.
[11] J. Sachs, P. Malaney, The economic and social burden of malaria, Nature 415 (6872) (2002) 680–685.
[12] E.-E. Ooi, K.-T. Goh, D.J. Gubler, Dengue prevention and 35 years of vector control in Singapore, Emerg. Infect. Dis. 12 (6) (2006) 887–893.
[13] M. Reidy, F. Ryan, D. Hogan, S. Lacey, C. Buckley, Preparedness of hospitals in the Republic of Ireland for an influenza pandemic, an infection control perspective, BMC Public Health 15 (2015) 847, http://dx.doi.org/10.1186/s12889-015-2025-6.

[14] T. Rebmann, APIC state-of-the-art report: the role of the infection preventionist in emergency management, Am. J. Infect. Control 37 (4) (2009) 271–281.

[15] M. Jovanovic, S. Radovanovic, M. Vukicevic, S. Van Poucke, B. Delibasic, Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression, Artif. Intell. Med. 72 (2016) 12–21.

[16] M.E. Beatty, A. Stone, D.W. Fitzsimons, J.N. Hanna, S.K. Lam, S. Vong, M.G. Guzman, J.F. Mendez-Galvan, S.B. Halstead, G.W. Letson, J. Kuritsky, R. Mahoney, H.S. Margolisfor T.A.-P. and A.D.P.B.S.W. Group, Best practices in dengue surveillance: a report from the Asia-Pacific and Americas dengue prevention boards, PLoS Negl. Trop. Dis. 4 (2010) e890, http://dx.doi.org/10.1371/journal.pntd.0000890.

[17] S.V. Poucke, Z. Zhang, M. Schmitz, M. Vukicevic, M.V. Laenen, L.A. Celi, et al., Scalable predictive analysis in critically Ill patients using a visual open data analysis platform, PLOS ONE 11 (1) (2016) e0145791.

[18] S. Van Poucke, M. Thomeer, J. Heath, M. Vukicevic, Are randomized controlled trials the (g)old standard? From clinical intelligence to prescriptive analytics, J. Med. Internet Res. 18 (7) (2016) e185.

[19] Y. Zhang, T. Wang, K. Liu, Y. Xia, Y. Lu, Q. Jing, et al., Developing a time series predictive model for Dengue in Zhongshan, China based on weather and Guangzhou Dengue surveillance data, PLoS Negl. Trop. Dis. 10 (2) (2016) e0004473 (S.V. Scarpino (Ed.)).

[20] M.A. Johansson, N.G. Reich, A. Hota, J.S. Brownstein, M. Santillana, Evaluating the performance of infectious disease forecasts: a comparison of climate-driven and seasonal dengue forecasts for Mexico, Sci. Rep. 26 (6) (2016) 33707.

[21] X. Zhang, F. Hou, Z. Qiao, X. Li, L. Zhou, Y. Liu, et al., Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014, BMJ Open. 6 (10) (2016) e011038.

[22] B. Pell, Y. Kuang, C. Viboud, G. Chowell, Using phenomenological models for forecasting the 2015 Ebola challenge, Epidemics (2016), http://dx.doi.org/10.1016/j.epidem.2016.11.002.

[23] L. Held, S. Meyer, J. Bracher, Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture: L. HELD, S. MEYER AND J. BRACHER, Stat. Med. (2017), http://dx.doi.org/10.1002/sim.7363.

[24] S. Funk, A. Camacho, A.J. Kucharski, R.M. Eggo, W.J. Edmunds, Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model, Epidemics (2016), http://dx.doi.org/10.1016/j.epidem.2016.11.003.

[25] A.I. Cotar, E. Falcuta, L.F. Prioteasa, S. Dinu, C.S. Ceianu, S. Paz, Transmission dynamics of the West Nile virus in mosquito vector populations under the influence of weather factors in the danube delta, Romania, EcoHealth (2016).

[26] N. Sundell, L.-M. Andersson, R. Brittain-Long, M. Lindh, J. Westin, A four year seasonal survey of the relationship between outdoor climate and epidemiology of viral respiratory tract infections in a temperate climate, J. Clin. Virol. Off. Publ. Pan. Am. Soc. Clin. Virol. 7 (84) (2016) 59–63.

[27] M.K. Butterworth, C.W. Morin, A.C. Comrie, An analysis of the potential impact of climate change on dengue transmission in the Southeastern United States, Environ. Health Perspect. (2016).

[28] P. Wang, W.B. Goggins, E.Y.Y. Chan, Hand, foot and mouth disease in Hong Kong: a time-series analysis on its relationship with weather, PloS One 11 (8) (2016) e0161006.

[29] E.R. Deyle, M.C. Maher, R.D. Hernandez, S. Basu, G. Sugihara, Global environmental drivers of influenza, Proc. Natl. Acad. Sci. USA (2016).

[30] A.C. Lowen, S. Mubareka, J. Steel, P. Palese, Influenza virus transmission is dependent on relative humidity and temperature, PLoS Pathog. 3 (10) (2007) e151.

[31] J. Shaman, V.E. Pitzer, C. Viboud, B.T. Grenfell, M. Lipsitch, Absolute humidity and the seasonal onset of influenza in the continental United States, PLoS Biol. 8 (2) (2010) e1000316 (N.M. Ferguson (Ed.)).

[32] D.A. da Cruz Ferreira, C.M. Degener, C. de Almeida Marques-Toledo, M.M. Bendati, L.O. Fetzer, C.P. Teixeira, Á.E. Eiras, Meteorological variables and mosquito monitoring are good predictors for infestation trends of Aedes aegypti, the vector of dengue, chikungunya and Zika, Parasit. Vectors 10 (2017), http://dx.doi.org/10.1186/s13071-017-2025-8.

[33] E.A. Mordecai, J.M. Cohen, M.V. Evans, P. Gudapati, L.R. Johnson, C.A. Lippi, et al., Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models, PLoS Negl. Trop. Dis. 11 (4) (2017) e0005568 (B. Althouse (Ed.)).

[34] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. 58 (1) (1996) 267–288.

[35] A. Albert, J.A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, Biometrika 71 (1) (1984) 1–10.

[36] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.

[37] World Health Organization, Pandemic Influenza Severity Assessment (PISA): A WHO Guide to Assess the Severity of Influenza in Seasonal Epidemics and Pandemics [Internet], World Health Organization, 2017. Available from: < http://www.who.int/iris/handle/10665/259392 > .

[38] Y. Shi, X. Liu, S.-Y. Kok, J. Rajarethinam, S. Liang, G. Yap, C.-S. Chong, K.-S. Lee, S.S.Y. Tan, C.K.Y. Chin, A. Lo, W. Kong, L.C. Ng, A.R. Cook, Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in Singapore, Environ. Health Perspect. 124 (2016) 1369–1375, http://dx.doi.org/10.1289/ehp.1509981.

[39] M.C. Peel, B.L. Finlayson, T.A. McMahon, Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci. 11 (5) (2007) 1633–1644.

[40] T.-W. Chuang, L.F. Chaves, P.-J. Chen, Effects of local and regional climatic fluctuations on dengue outbreaks in southern Taiwan, PLOS ONE 12 (6) (2017) e0178698 (A.S. Azman (Ed.)).

[41] E.K. Kipruto, A.O. Ochieng, D.N. Anyona, M. Mbalanya, E.N. Mutua, D. Onguru, I.K. Nyamongo, B.B.A. Estambale, Effect of climatic variability on malaria trends in Baringo County, Kenya, Malar. J. 16 (2017), http://dx.doi.org/10.1186/s12936-017-1848-2.

[42] C. Wang, K. Cao, Y. Zhang, L. Fang, X. Li, Q. Xu, F. Huang, L. Tao, J. Guo, Q. Gao, X. Guo, Different effects of meteorological factors on hand, foot and mouth disease in various climates: a spatial panel data model analysis, BMC Infect. Dis. 16 (2016), http://dx.doi.org/10.1186/s12879-016-1560-9.

[43] K. Harigane, A. Sumi, K. Mise, N. Kobayashi, The role of temperature in reported chickenpox cases from 2000 to 2011 in Japan, Epidemiol. Infect. 143 (12) (2015) 2666–2678.

[44] National Institute of Infectious Diseases, Japan [Internet] (cited 2016 Jan 19). Available from: < http://www0.nih.go.jp/niid/index-e.html > .

[45] Bureau of Epidemiology, Thailand [Internet] (cited 2016 Jan 19). Available from: < http://203.157.15.110/boeeng/ > .

[46] Bureau of Epidemiology, DDC, MOPH, National Disease Surveillance (Report 506) [Internet] (cited 2016 Nov 7). Available from: < http://www.boe.moph.go.th/boedb/surdata/ > .

[47] Weekly Infectious Diseases Bulletin|Ministry of Health [Internet] (cited 2016 Jan 20). Available from: < https://www.moh.gov.sg/content/moh_web/home/statistics/infectiousDiseasesStatistics/weekly_infectiousdiseasesbulletin.html > .

[48] Taiwan National Infectious Disease Statistics System [Internet] (cited 2016 Jan 26). Available from: < http://nidss.cdc.gov.tw/en/SingleDisease.aspx?dc=1&dt=4&disease=061&position=1 > .

[49] X. Zhao, EpiWeek: Conversion between Epidemiological Weeks and Calendar Dates [Internet] (2016) (cited 2016 Dec 27). Available from: < https://cran.r-project.org/web/packages/EpiWeek/index.html > .

[50] Weather Forecast & Reports – Long Range & Local | Wunderground | Weather Underground [Internet] (cited 2016 Jan 20). Available from: < http://www.wunderground.com/ > .

[51] Japan Meteorological Agency [Internet] (cited 2016 Jan 20). Available from: < http://www.jma.go.jp/jma/indexe.html > .

[52] B.T. Grenfell, O.N. Bjørnstad, J. Kappey, Travelling waves and spatial hierarchies in measles epidemics, Nature 414 (6865) (2001 13) 716–723.

[53] M. Farge, Wavelet transforms and their applications to turbulence, Annu. Rev. Fluid Mech. 24 (1) (1992) 395–458.

[54] C. Torrence, G.P. Compo, A practical guide to wavelet analysis, Bull. Am. Meteorol. Soc. 79 (1) (1998) 61–78.

[55] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014. Available from: < http://www.R-project.org/ > .

[56] Y. Chen, A.R. Cook, A.X.L. Lim, Randomness of dengue outbreaks on the equator, Emerg. Infect. Dis. 21 (9) (2015) 1651–1653.

[57] H. Feng, G. Duan, R. Zhang, W. Zhang, Time series analysis of hand-foot-mouth disease hospitalization in Zhengzhou: establishment of forecasting models using climate variables as predictors, PLoS ONE 9 (1) (2014) e87916 (B.J. Cowling (Ed.)).

[58] D. Zhao, L. Wang, J. Cheng, J. Xu, Z. Xu, M. Xie, et al., Impact of weather factors on hand, foot and mouth disease, and its role in short-term incidence trend forecast in Huainan City, Anhui Province. Int J Biometeorol (2016).

[59] X. Song, J. Xiao, J. Deng, Q. Kang, Y. Zhang, J. Xu, Time series analysis of influenza incidence in Chinese provinces from 2004 to 2011, Medicine (Baltimore) 95 (26) (2016) e3929.

[60] P.F. Craigmile, N. Kim, S.A. Fernandez, B.K. Bonsu, Modeling and detection of respiratory-related outbreak signatures, BMC Med. Inf. Decis. Making 5 (7) (2007) 28.

[61] F.F. Nobre, A.B. Monteiro, P.R. Telles, G.D. Williamson, Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology, Stat. Med. 20 (20) (2001) 3051–3069.

[62] E. Goldstein, B.J. Cowling, A.E. Aiello, S. Takahashi, G. King, Y. Lu, et al., Estimating incidence curves of several infections using symptom surveillance data, PLoS ONE 6 (8) (2011) e23380 (A. Yates (Ed.)).

[63] X.-S. Zhang, R. Pebody, A. Charlett, D. de Angelis, P. Birrell, H. Kang, M. Baguelin, Y.H. Choi, Estimating and modelling the transmissibility of middle east respiratory syndrome corona virus during the 2015 outbreak in the Republic of Korea, Influenza Other Respir Viruses 11 (2017) 434–444, http://dx.doi.org/10.1111/irv.12467.

[64] X. Zhao, V.J. Fang, S.E. Ohmit, A.S. Monto, A.R. Cook, B.J. Cowling, Quantifying protection against influenza virus infection measured by hemagglutination-inhibition assays in vaccine trials, Epidemiology 27 (1) (2016) 143–151.

[65] E.J. Kasowski, R.J. Garten, C.B. Bridges, Influenza pandemic epidemiologic and virologic diversity: reminding ourselves of the possibilities, Clin. Infect. Dis. 52 (Suppl. 1) (2011) S44–S49.

[66] M.J. Keeling, K.T.D. Eames, Networks and epidemic models, J. R. Soc. Interface 2 (4) (2005) 295–307.

[67] D.L. Chao, M.E. Halloran, V.J. Obenchain Jr, I.M.L. FluTE, A publicly available stochastic influenza epidemic simulation model, PLOS Comput. Biol. 6 (1) (2010) e1000656.

[68] World Health Organization, New Influenza A (H1N1) Virus: Global Epidemiological Situation, June 2009, 2009. Available from: < http://www.who.int/iris/handle/

10665/241366 > .
[69] M. Enserink, INFECTIOUS DISEASES: massive outbreak draws fresh attention to little-known virus, Science 311 (5764) (2006) 1085a–1085a.
[70] G.S. Campos, A.C. Bandeira, S.I. Sardi, Zika virus outbreak, Bahia, Brazil, Emerg Infect Dis. 21 (10) (2015) 1885–1886.
[71] N. Halder, J.K. Kelso, G.J. Milne, Analysis of the effectiveness of interventions used during the 2009 A/H1N1 influenza pandemic, BMC Public Health 10 (2010) 168, http://dx.doi.org/10.1186/1471-2458-10-168.
[72] B.L. Dickens, J. Yang, A.R. Cook, L.R. Carrasco, Time to empower release of insects carrying a dominant lethal and *Wolbachia* against Zika, Open Forum Infect Dis 3 (2) (2016) ofw103.
[73] J. Páez Chávez, T. Götz, S. Siegmund, K.P. Wijaya, An SIR-Dengue transmission model with seasonal effects and impulsive control, Math. Biosci. 1 (289) (2017) 29–39.