BMC Medical Research Methodology

RESEARCH Open Access

SpatialWavePredict: a tutorial-based primer and toolbox for forecasting growth trajectories using the ensemble spatial wave sub-epidemic modeling framework

Gerardo Chowell[1,2*], Amna Tariq[3], Sushma Dahal[1], Amanda Bleichrodt[1], Ruiyan Luo[1] and James M. Hyman[4]

Abstract

Background Dynamical mathematical models defined by a system of differential equations are typically not easily accessible to non-experts. However, forecasts based on these types of models can help gain insights into the mechanisms driving the process and may outcompete simpler phenomenological growth models. Here we introduce a friendly toolbox, *SpatialWavePredict*, to characterize and forecast the spatial wave sub-epidemic model, which captures diverse wave dynamics by aggregating multiple asynchronous growth processes and has outperformed simpler phenomenological growth models in short-term forecasts of various infectious diseases outbreaks including SARS, Ebola, and the early waves of the COVID-19 pandemic in the US.

Results This tutorial-based primer introduces and illustrates a user-friendly MATLAB toolbox for fitting and forecasting time-series trajectories using an ensemble spatial wave sub-epidemic model based on ordinary differential equations. Scientists, policymakers, and students can use the toolbox to conduct real-time short-term forecasts. The five-parameter epidemic wave model in the toolbox aggregates linked overlapping sub-epidemics and captures a rich spectrum of epidemic wave dynamics, including oscillatory wave behavior and plateaus. An ensemble strategy aims to improve forecasting performance by combining the resulting top-ranked models. The toolbox provides a tutorial for forecasting time-series trajectories, including the full uncertainty distribution derived through parametric bootstrapping, which is needed to construct prediction intervals and evaluate their accuracy. Functions are available to assess forecasting performance, estimation methods, error structures in the data, and forecasting horizons. The toolbox also includes functions to quantify forecasting performance using metrics that evaluate point and distributional forecasts, including the weighted interval score.

Conclusions We have developed the first comprehensive toolbox to characterize and forecast time-series data using an ensemble spatial wave sub-epidemic wave model. As an epidemic situation or contagion occurs, the tools presented in this tutorial can facilitate policymakers to guide the implementation of containment strategies and assess the impact of control interventions. We demonstrate the functionality of the toolbox with examples, including a tutorial video, and is illustrated using daily data on the COVID-19 pandemic in the USA.

Keywords MATLAB toolbox, Real-time forecasting, Dynamic growth model, Spatial wave sub-epidemic wave model, Ensemble model, Complex epidemic patterns

*Correspondence:
Gerardo Chowell
gchowell@gsu.edu
Full list of author information is available at the end of the article

Background

Developing reliable methods for forecasting dynamic growth processes is critical for decision-making in problems ranging from predicting the weather, forecasting the trajectory of an emerging epidemic, the growth or decline of economic variables, election outcomes, and sporting events [1]. While statistical methods such as ARIMA and exponential smoothing are robust and broadly competitive for forecasting time series [2–6], dynamical mathematical models defined by a system of differential equations are typically not easily accessible to non-experts. However, forecasts based on these types of models can help characterize the mechanisms driving the process [7]. They may offer higher forecasting performance than purely statistical approaches based on statistical evaluation criteria like mean absolute and squared errors [8–11]. Here we focus on dynamical models that can characterize growth processes that give rise to waves of variable shapes and sizes [12–14]. The complexity of this family of growth models ranges from single differential equation models with a few parameters, such as the 3-parameter generalized-logistic growth model (GLM) [14], to systems of ordinary differential equations (ODEs) that capture diverse wave dynamics by aggregating multiple asynchronous growth processes [13]. The spatial wave sub-epidemic framework has outperformed simpler phenomenological growth models in forecasts of various infectious diseases, including severe acute respiratory syndrome (SARS), Ebola, and the early waves of the coronavirus disease 2019 (COVID-19) pandemic in the United States (US) [13, 15].

This tutorial paper introduces a user-friendly MATLAB toolbox to fit and forecast time-series trajectories using the spatial wave sub-epidemic dynamic growth model based on ordinary differential equations, which was initially developed to characterize and derive short-term forecasts of epidemic trajectories [13, 16]. This mathematical framework characterizes time-series trajectories by aggregating multiple asynchronous growth processes. Each growth process (i.e., sub-epidemic) is modeled using a simple phenomenological growth model such as the generalized logistic growth model (GLM). This framework supports a family of growth models that yield similar fits to the calibration data, but their corresponding forecasts could produce diverse trajectories. Hence, we also incorporate ensemble techniques to combine the resulting models to boost forecasting performance [16, 17].

This toolbox is written for a diverse audience, including students training in time-series forecasting. It allows the user to conduct parameter estimation and forecasting with quantified uncertainty and evaluate forecasting performance using a set of standard metrics, including the coverage of the 95% prediction interval and the weighted interval score, which account for the uncertainty of the predictions. The toolbox allows scientists and policymakers to generate short-term forecasts by relying on minimal data of the process of interest, such as an unfolding epidemic or natural disaster.

The toolbox provides prediction intervals and allows the user to employ different estimation methods, assumptions of the error structure, and forecasting horizons. For instance, the toolbox includes estimation methods such as the nonlinear least squares estimation and maximum likelihood estimation (MLE) with different assumptions about the error structure of the observed data, including Poisson, negative binomial, and normal distributions, as well as quantification of the uncertainty based on a parametric bootstrapping approach. The model also provides flexibility to choose the underlying building block of the growth process. In addition, the toolbox includes functions to derive weighted and unweighted ensembles based on the resulting top-ranked models. The full functionality of the toolbox is illustrated using daily time series of COVID-19 cases in the US, and in the process, shows that this framework outcompetes simpler single growth models and simple time-series models (e.g., ARIMA, GAM, SLR) in calibration and forecasting performance.

We start by describing the format of the input time-series data, followed by the methods employed for parameter estimation. Next, we describe the underlying methodology, user parameters, and functions to calibrate, evaluate, and display the model fits. Finally, we introduce the functions to generate, display, and quantify the performance of model-based forecasts with specific examples in the context of the daily COVID-19 case data reported in the USA. A tutorial video that demonstrates the toolbox functionality is available at: https://www.youtube.com/watch?v=qxuF_tTzcR8&t=47s.

Implementation

In this section, we describe the methods implemented in this toolbox and provide a brief overview of the toolbox functions.

Installing the toolbox

- Download the MATLAB code located in the folder ***spatialWave_subepidemicFramework code*** from the GitHub repository: https://github.com/gchowell/spatial_wave_subepidemic_framework.
- Create an 'input' folder in your working directory where your input data will be stored.
- Create an 'output' folder in your working directory where the output files will be stored.

- Open a MATLAB session.

Overview of the toolbox functions

The methodological workflow of the tutorial is organized as follows: (1) plotting model simulations, (2) fitting the models to data with quantified uncertainty, (3) plotting the resulting model fits and calibration performance metrics, and (4) plotting model-based forecasts and the associated forecasting performance metrics. Table 1 and Supplementary Table 1 list the names of both user and internal functions associated with the toolbox, along with a brief description of their role. As described below, the user needs to specify the parameters related to model fitting and forecasting in the default `options_fit.m` and `options_forecast.m files`.

Parameter estimation method

Let $f(t,\Theta)$ denote the expected curve of the epidemic's trajectory. We can estimate model parameters Θ by fitting the model solution to the observed data via nonlinear least squares [18] or maximum likelihood estimation with specific assumptions about the error structure in the data [19] by specifying parameter`<method1>`in the `options.m` file. For nonlinear least squares (i.e., `<method1>=0`), this is achieved by searching for the set of parameters $\widehat{\Theta}$ that minimizes the sum of squared differences between the smoothed data $y_{t_j}=y_{t_1},y_{t_2}\ldots y_{t_{n_d}}$ and the model mean, corresponding to $f(t,\Theta)$. That is, $\Theta=(C_{thr},r,p,q,K_0)$ in the sub-epidemic wave model (given below) is estimated by $\widehat{\Theta}=\text{argmin}\sum_{j=1}^{n_d}\left(f\left(t_j,\Theta\right)-y_{t_j}\right)^2$. We estimate the parameter C_{thr} through simple discretization of its range of plausible values. Our estimation procedure consists of two steps. First, for each C_{thr}, we search for the set of parameters (r,p,q,K_0) that yield the best fit to the data. Then we choose $C_{thr}$and the corresponding estimates of other parameters leading to the overall best-fit to the data.

Nonlinear least squares estimation weighs each of the data points equally and does not explicitly require a specific distributional assumption for y_t, except for the first moment $E[y_t]=f(t_i;\Theta)$. That is, the mean of the observed data at time t is equivalent to the expected count denoted by $f(t,\Theta)$at time t [20]. This method yields asymptotically unbiased point estimates regardless of any misspecification of the variance-covariance error structure. Hence, the estimated model mean $f(t_i,\widehat{\Theta})$ yields the best fit to observed data y_{t_i}in terms of squared L2 norm. We can solve the nonlinear least squares optimization problem using the *fmincon* function in MATLAB. Moreover, we also employ MATLAB's MultiStart feature to specify the number of random initial guesses of the model parameters using the parameter `<numstart-points>` in the `options.m` file in order to search thoroughly for a global minimum, check that the solution is unique, and the parameters are identifiable.

We can also estimate parameters via maximum likelihood estimation (MLE) [19] and assume different error structures in the data (e.g., Poisson, negative binomial). The log-likelihood expressions derived for different error structures are specified below.

Poisson

For a Poisson error structure, the full log-likelihood of Poisson (i.e., `<method1>=1`) is given by:

Table 1 Description of the user functions available in the *SpatialWavePredict* toolbox

Function	Role
`options.m`	Specifies the parameters related to model fitting, including the characteristics of the time series data, the sub-epidemic model, parameter estimation method, error structure, smoothing, and calibration period. The structure of the `options.m` file is given in **Supplementary Text** 1.
`options_forecast.m`	Specifies the parameters related to the forecast, including the forecasting period, the type of ensemble weight for the ensemble models, and whether the forecasts will be evaluated. The structure of the `options_forecast.m` file is given in **Supplementary Text** 2.
`plot_SW_subepidemic.m`	Plots simulations of the spatial wave sub-epidemic model.
`Run_SW_subepidemicFramework.m`	Derives the top-ranking sub-epidemic wave models to data with quantified uncertainty.
`plotRankings_SW_subepidemicFramework.m`	Plots the mean model fits of the top-ranking models, including their sub-epidemic profiles, and the associated quality of model fit metrics, including the AICc, the relative likelihood, and the evidence ratio.
`plotFit_SW_subepidemicFramework.m`	Displays the model fit and 95% prediction interval, as well as the empirical distribution of the parameters. It also saves output .csv files in the output folder with the model fit, the parameter estimates, including 95% CIs, and the calibration performance metrics.
`plotForecast_SW_subepidemicFramework.m`	Displays the model-based forecast and the performance metrics of the forecast. Moreover, the data associated with the forecasts, the parameter estimates, as well as the calibration and forecasting performance metrics are saved as .csv files in the output folder.

$$\sum_{i=1}^{n}\left\{y_i ln(\mu_i) - ln(y_i!) - \mu_i\right\},$$

where $\mu_i = f(t_i,\theta)$ denotes the mean of $y_i$at time t_i. The number of parameters is just the number of parameters estimated in the dynamical model based on ordinary differential equations.

Negative binomial

Let $r > 0$ denote the number of failures until the experiment is stopped, $p \in [0,1]$ denote the success probability in each experiment. The number of successes y before the r-th failure occurs has a **negative binomial distribution** given by:

$$f(y|r,p) = \binom{r+y-1}{y} p^y(1-p)^r = \frac{1}{y!}\prod_{j=0}^{y-1}(j+r).p^y(1-p)^r$$

with mean $= \mu = \frac{rp}{(1-p)}$, variance $= \sigma^2 = \frac{rp}{(1-p)^2} > \mu$. For n observations $y_1,\ldots,y_n$, the full log-likelihood is

$$l(r,p) = \sum_{i=1}^{n}\left\{\left\{\sum_{j=0}^{y_i-1}\ln(j+r)\right\} + y_i\ln(p_i) + rln(1-p_i) - \ln(y_i!)\right\}, \tag{1.1}$$

which can be expressed with μ and σ^2 by plugging-in $p = 1 - \frac{\mu}{\sigma^2}$ and $r = \frac{\mu^2}{\sigma^2-\mu}$.

There are different types of variances commonly used in a negative binomial distribution. If the variance scales linearly with the mean: $\sigma^2 = \mu + \alpha\mu$, (i.e., <method1>=3 in options.m), then $p = \frac{\alpha}{1+\alpha}$ and $r = \mu/\alpha$. Let $\mu = f(t,\theta)$be the mean curve to be estimated from the differential equation. The full log-likelihood (1.1) can be expressed as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{n}\left\{\left\{\sum_{j=0}^{y_i-1} ln(j+\alpha^{-1}f(t_i,\theta))\right\} + y_i ln(\alpha) - (y_i+\alpha^{-1}f(t_i,\theta))ln(1+\alpha) - ln(y_i!)\right\}. \tag{1.2}$$

If the variance scales quadratically with the mean, $\sigma^2 = \mu + \alpha\mu^2$ (i.e., <method1>=4 in options.m), then $p = \frac{\alpha\mu}{1+\alpha\mu}$and $r = 1/\alpha$. The full log-likelihood (1.1) can be expressed as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{n}\left\{\left\{\sum_{j=0}^{y_i-1} ln(j+\alpha^{-1})\right\} + y_i ln\left(\alpha f(t_i,\theta)\right) - (y_i+\alpha^{-1})ln(1+\alpha f(t_i,\theta)) - ln(y_i!)\right\}. \tag{1.3}$$

The more general form of variance is $\sigma^2 = \mu + \alpha\mu^d$ (i.e., <method1>=5 in options.m) with any $-\infty < d < \infty$. Then the full log-likelihood (1.1) can be expressed as follows:

$$l(\theta,\alpha) = \sum_{i=1}^{n}\left[\left\{\sum_{j=0}^{y_i-1} ln(j+\alpha^{-1}{\mu_i}^{2-d})\right\} + y_i ln\left(\alpha{\mu_i}^{d-1}\right) - (y_i+\alpha^{-1}{\mu_i}^{2-d})ln(1+\alpha{\mu_i}^{d-1}) - ln(y_i!)\right], \tag{1.4}$$

where $\mu_i = f(t_i,\theta)$.

The number of parameters is 1 plus the number of parameters in the dynamical model based on ordinary differential equations (ODE) for (1.2) ~ (1.3), and 2 plus the number of parameters in the dynamical model for (1.4) if d is also estimated via MLE. Assuming Poisson or negative binomial error structures in the data, we can estimate parameters using MLE by specifying parameters in the options.m file, such as <method1>=1 & <dist1>=1 for Poisson and <method1> & <dist1>=3, <method1>=4 & <dist1>=4, and <method1>=5 & <dist1>=5 for the different negative binomial error structures described above.

Parametric bootstrapping

To quantify parameter uncertainty, we follow a parametric bootstrapping approach which allows the computation of standard errors and related statistics in the absence of closed-form formulas [21]. We generate B bootstrap samples from the best-fit model $f(t,\widehat{\Theta})$, with an assumed error structure specified using parameter<dist1>in the options.m file to quantify the uncertainty of the parameter estimates and construct confidence intervals. Typically, the error structure in the data is modeled using a probability model such as the Poisson or negative binomial distribution. Using nonlinear least squares (`<method1>=0`), besides a normally distributed error structure (`<dist1>=0`), we can also assume a Poisson (`<dist1>=1`) or a negative binomial distribution (`<dist1>=2`) whereby the variance-to-mean ratio is empirically estimated from the time series. To estimate this constant ratio, we group a fixed number of observations (e.g., 7 observations for daily data into a bin across time), calculate the mean and variance for each bin, and then estimate a constant variance-to-mean ratio by calculating the average of the variance-to-mean ratios over these bins.

Using the best-fit model $f(t,\widehat{\Theta})$, we generate B-times replicated simulated datasets of size n_d, where the observation at time $t_j$is sampled from the corresponding distribution specified by `<dist1>`. Next, we refit the model to each of the B simulated datasets to re-estimate the parameters using the same estimation method for the bootstrap sample as for the original data. This allows us to quantify the uncertainty of the estimate using that method. The new parameter estimates for each realization are denoted by $\widehat{\Theta}_b$, where $b = 1,2,\ldots,B$. Using the sets of re-estimated parameters ($\widehat{\Theta}_b$), it is possible to characterize the empirical distribution of each estimate, calculate the variance, and construct confidence intervals for each parameter. The resulting uncertainty around the model fit can similarly be obtained from $f\left(t,\widehat{\Theta}_1\right)$, $f\left(t,\widehat{\Theta}_2\right),\ldots,f(t,\widehat{\Theta}_B)$. We characterize the uncertainty using 300 bootstrap realizations (i.e., parameter `<B>=300` in the `options.m` file).

Model-based forecasts with quantified uncertainty

Forecasting the model $f\left(t,\widehat{\Theta}\right)$, h days ahead is based on the estimate $f(t+h,\widehat{\Theta})$. The uncertainty of the forecasted value can be obtained using the previously described parametric bootstrap method. Let

$$f\left(t+h,\widehat{\Theta}_1\right), f\left(t+h,\widehat{\Theta}_2\right),\ldots,f(t+h,\widehat{\Theta}_B)$$

denote the forecasted value of the current state of the system propagated by a horizon of h time units, where $\widehat{\Theta}_b$ denotes the estimation of parameter set Θ from the b_{th} bootstrap sample. We can use these values to calculate the bootstrap variance to measure the uncertainty of the forecasts and use the 2.5% and 97.5% percentiles to construct the 95% prediction intervals (95% PIs). We can set the forecasting horizon using the parameter `<forecastingperiod1>` in the `options_forecast.m` file. The structure of the `options_forecast.m` file is described in Supplementary Text 2.

For the COVID-19 case data employed for illustration purposes, we fit the models by the nonlinear least squares method assuming a normal error structure (i.e., `<method1>=0` and `<dist1>=0`) (Fig. 1).

Sub-epidemic wave model

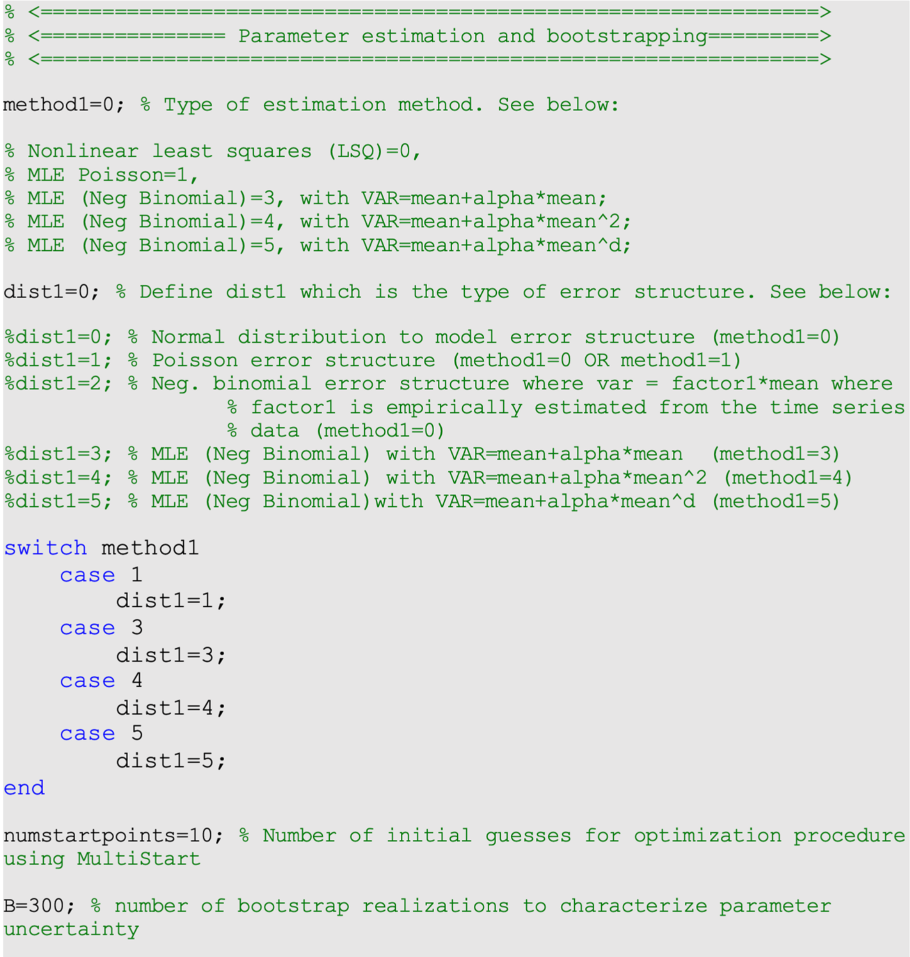
We use a spatial wave model with up to 5 parameters that aggregate linked overlapping sub-epidemics [13]. This sub-epidemic framework can characterize diverse epidemic patterns, including the epidemic plateaus, where the epidemic stabilizes at a high level for an extended period and the epidemic waves have multiple peaks. The strength (e.g., weak vs. strong) of their overlap determines when the next sub-epidemic is triggered and is controlled by the onset threshold parameter, C_{thr}. The mathematical equation for the sub-epidemic building block is the 3-parameter generalized-logistic growth model (GLM), which is specified by setting the parameter `<flag1>=1` in the `options.m` file. This growth model has performed well in short-term forecasts of single outbreak trajectories for different infectious diseases, including COVID-19 [22–24]. Alternative growth equations to model the sub-epidemic building block include the 3-parameter Richards model (`<flag1>=4`) and the 2-parameter logistic growth model (`<flag1>=2`). The following differential equation gives the generalized-logistic growth model (GLM):

$$\frac{dC(t)}{dt} = C'(t) = rC^p(t)\left(1 - \frac{C(t)}{K_0}\right),$$

where $C(t)$ denotes the cumulative curve at time t, and $\frac{dC(t)}{dt}$ describes the epidemic's incidence curve over time t. The positive parameter r denotes the growth rate per unit of time, K_0 is the final outbreak size, and $p \in [0,1]$ is the "scaling of growth" parameter which allows the model to capture early sub-exponential and exponential growth patterns. If $p = 0$, this equation describes a constant incidence over time, while $p = 1$ indicates that the early growth phase is exponential. Intermediate values of $p(0 < p < 1)$ describe early sub-exponential (e.g., polynomial) growth dynamics. The sub-epidemic wave model consists of a system of coupled differential equations:

$$\frac{dC_i(t)}{dt} = rA_{i-1}(t)C_i(t)^p\left(1 - \frac{C_i(t)}{K_i}\right).$$

Here, $C_i(t)$ is the cumulative number of infections for sub-epidemic i, and K_i is the size of the i_{th} sub-epidemic where $i = 1,\ldots,n$. Starting from an initial sub-epidemic size K_0, the size of consecutive sub-epidemics K_i decline



```
% <===============================================================>
% <=============== Parameter estimation and bootstrapping=========>
% <===============================================================>

method1=0; % Type of estimation method. See below:

% Nonlinear least squares (LSQ)=0,
% MLE Poisson=1,
% MLE (Neg Binomial)=3, with VAR=mean+alpha*mean;
% MLE (Neg Binomial)=4, with VAR=mean+alpha*mean^2;
% MLE (Neg Binomial)=5, with VAR=mean+alpha*mean^d;

dist1=0; % Define dist1 which is the type of error structure. See below:

%dist1=0; % Normal distribution to model error structure (method1=0)
%dist1=1; % Poisson error structure (method1=0 OR method1=1)
%dist1=2; % Neg. binomial error structure where var = factor1*mean where
                  % factor1 is empirically estimated from the time series
                  % data (method1=0)
%dist1=3; % MLE (Neg Binomial) with VAR=mean+alpha*mean  (method1=3)
%dist1=4; % MLE (Neg Binomial) with VAR=mean+alpha*mean^2 (method1=4)
%dist1=5; % MLE (Neg Binomial)with VAR=mean+alpha*mean^d (method1=5)

switch method1
    case 1
        dist1=1;
    case 3
        dist1=3;
    case 4
        dist1=4;
    case 5
        dist1=5;
end

numstartpoints=10; % Number of initial guesses for optimization procedure
using MultiStart

B=300; % number of bootstrap realizations to characterize parameter
uncertainty
```

Fig. 1 Contents of `options.m` file, the values of the parameters related to the parameter estimation method and parametric bootstrapping

at the rate q following an exponential or power-law function as described below. Hence, a total of 5 parameters (r, p, C_{thr}, q, K_0) for $i = 1, \ldots, n$ are needed to characterize a sub-epidemic wave composed of two or more sub-epidemics.

The onset timing of the subsequent $(i+1)_{th}$ sub-epidemic is determined by the indicator variable $A_i(t)$. This results in a coupled system of sub-epidemics where the $(i+1)_{th}$ sub-epidemic is triggered when the cumulative curve for the i_{th} sub-epidemic exceeds a total of C_{thr}. The sub-epidemics *overlap* because the $(i+1)_{th}$sub-epidemic takes off before the i_{th} sub-epidemic completes its course. That is,

$$A_i(t) = \begin{cases} 1 & C_i(t) > C_{thr} \\ 0 & Otherwise \end{cases}, i = 1, 2, \ldots, n-1.$$

The threshold parameters are defined so that $1 \leq C_{thr} < K_0$ and $A_0(t) = 1$ for the first sub-epidemic. The maximum number of sub-epidemics considered in the epidemic wave trajectory is specified using parameter `<npatches_fixed>` in the `options.m` file. Here, we set `<npatches_fixed>=3`. The initial number of cases is given by $C_1(0) = I_0$, where I_0 is the initial number of cases in the observed data.

In this framework, the size of the subsequent i_{th} sub-epidemic (K_i) remains steady or declines due to the effects of behavior changes or interventions. We consider both exponential and inverse decline functions to model the size of consecutive sub-epidemics described below.

Exponential decline of sub-epidemic sizes

If consecutive sub-epidemics follow exponential decline, then K_i is given by:

$$K_i = K_0 e^{-q(i-1)},$$

where K_0 is the size of the initial sub-epidemic ($K_1 = K_0$). If $q = 0$, the model predicts an epidemic wave composed of sub-epidemics of the same size. When $q > 0$, the epidemic wave is composed of a finite number of sub-epidemics given by n_{tot} which is a function of $C_{thr}, q,$ andK_0 as follows:

$$n_{tot} = \left\lfloor -\frac{1}{q}\ln\left(\frac{C_{thr}}{K_0}\right) + 1 \right\rfloor.$$

where the brackets $\lfloor * \rfloor$ denote the largest integer that is smaller than or equal to $*$. The total size of the epidemic wave composed of n_{tot} overlapping sub-epidemics has the following closed-form solution:

$$K_{tot} = \sum_{i=1}^{n_{tot}} K_0 e^{-q(i-1)} = \frac{K_0(1 - e^{-qn_{tot}})}{1 - e^{-q}}.$$

The exponential sub-epidemic decline function can be selected by setting the parameter `<typedecline2>=1` in the `options.m` file.

Power-law decline of sub-epidemic sizes

If consecutive sub-epidemics decline according to the inverse function, we have:

$$K_i = K_0 \left(\frac{1}{i}\right)^q.$$

When q > 0, the total number of sub-epidemics n_{tot} comprising the epidemic wave is finite and given by:

$$n_{tot} = \left\lfloor \left(\frac{C_{thr}}{K_0}\right)^{-\frac{1}{q}} \right\rfloor.$$

The total size of an epidemic wave is given by the aggregation of n_{tot} overlapping sub-epidemics:

$$K_{tot} = \sum_{i=1}^{n_{tot}} K_0 \left(\frac{1}{i}\right)^q.$$

The power-law sub-epidemic decline function can be selected by setting the parameter `<typedecline2>=2` in the `options.m` file. Selecting the type of decline function that yields the best fit to the data is also possible by setting the parameter,

`<typedecline2>=[1 2]`.

Fixed sub-epidemic onset

We can also consider sub-epidemic wave models with a fixed onset time at 0. In this case, all sub-epidemics start at time 0, and the threshold parameter C_{thr} drops from the model. We use parameter `<onset_fixed>` in the `options.m` file to specify whether the onset timing of the sub-epidemics is fixed at time 0 (`<onset_fixed>=1`) or not (`<onset_fixed>=0`).

Top-ranked sub-epidemic models

To select the top-ranked sub-epidemic models, we analyze the Akaike information criterion (AIC_c) values of the set of best-fit sub-epidemic wave models with different values of C_{thr}. The AIC_c is given by [25, 26]:

$$AIC_c = -2log(likelihood) + 2m + \frac{2m(m+1)}{n_d - m - 1},$$

where m is the number of model parameters, and n_d is the number of data points. Specifically for normal distribution, the AIC_c is

$$AIC_c = n_d log(SSE) + 2m + \frac{2m(m+1)}{n_d - m - 1},$$

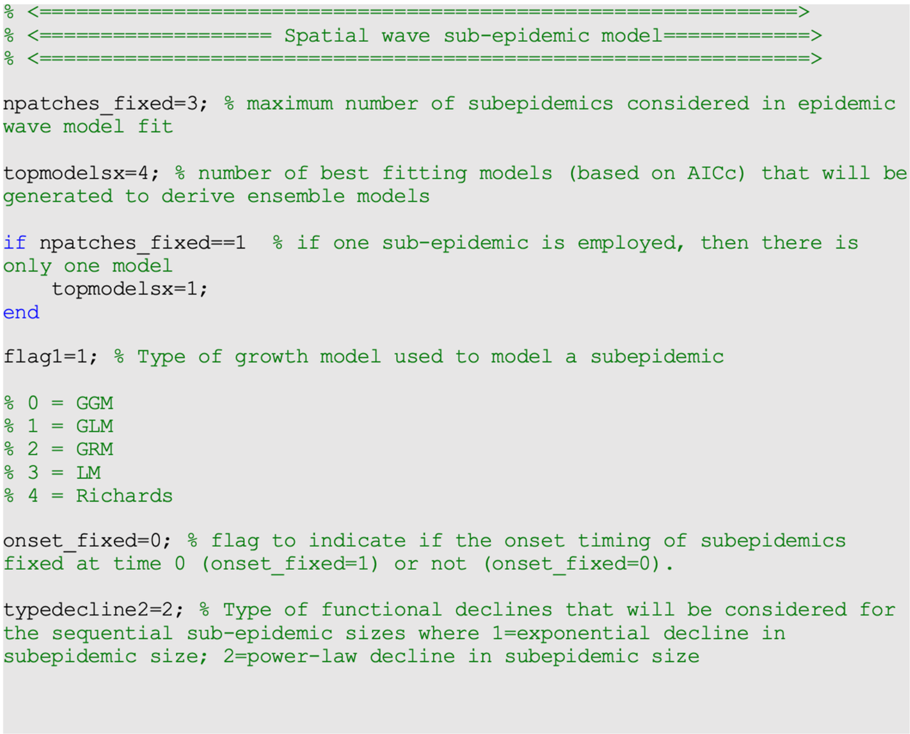
where $SSE = \sum_{j=1}^{n_d} (f\left(t_j, \widehat{\Theta}\right) - y_{t_j})^2$ is the sum of squared errors, m is the number of model parameters including parameter C_{thr}. Parameter `<topmodelsx>` in the `options.m` file is used to specify the number of top-ranked models that will be generated and used to derive ensemble models.

To illustrate the methodology, we set `<onset_fixed>=0`, `< typedecline2 >=2` (power-law decline) and analyzed four top-ranking sub-epidemic models (`<topmodelsx>=4`). The top-ranking models are used to construct three ensemble sub-epidemic models, which we refer to as: Ensemble(2), Ensemble(3), and Ensemble(4) (Fig. 2).

Plotting simulations of the spatial wave sub-epidemic model

Before fitting the growth model to the data, it is useful to check that the selected model yields simulations broadly consistent with the range of the time series data by generating model simulations with different parameter values. For example, if data show systematic differences that contrast with the model solutions, it may suggest that the model is not the best choice for the data at hand.

The function `plot_SW_subepidemic.m` can be used to plot model solutions where the user provides the type of growth model by passing parameter `<flag1>` (generalized-logistic growth model, Richards, Gompertz, etc.), the model parameter values, and the initial conditions as passing input parameters to the function in the following order: `<flag1>`, r, p, a, K, q, n, C_{thr},`<typedecline1>`, $C(0)$, and finally the duration of the simulation. For example, the following call plots a simulation of the spatial wave sub-epidemic model using as building block the generalized logistic growth model



```
% <==============================================================>
% <=================== Spatial wave sub-epidemic model============>
% <===============================================================>

npatches_fixed=3; % maximum number of subepidemics considered in epidemic
wave model fit

topmodelsx=4; % number of best fitting models (based on AICc) that will be
generated to derive ensemble models

if npatches_fixed==1  % if one sub-epidemic is employed, then there is
only one model
    topmodelsx=1;
end

flag1=1; % Type of growth model used to model a subepidemic

% 0 = GGM
% 1 = GLM
% 2 = GRM
% 3 = LM
% 4 = Richards

onset_fixed=0; % flag to indicate if the onset timing of subepidemics
fixed at time 0 (onset_fixed=1) or not (onset_fixed=0).

typedecline2=2; % Type of functional declines that will be considered for
the sequential sub-epidemic sizes where 1=exponential decline in
subepidemic size; 2=power-law decline in subepidemic size
```

Fig. 2 Contents of `options.m` file, the values of the parameters related to the sub-epidemic wave model and the number of top-ranked sub-epidemic wave models

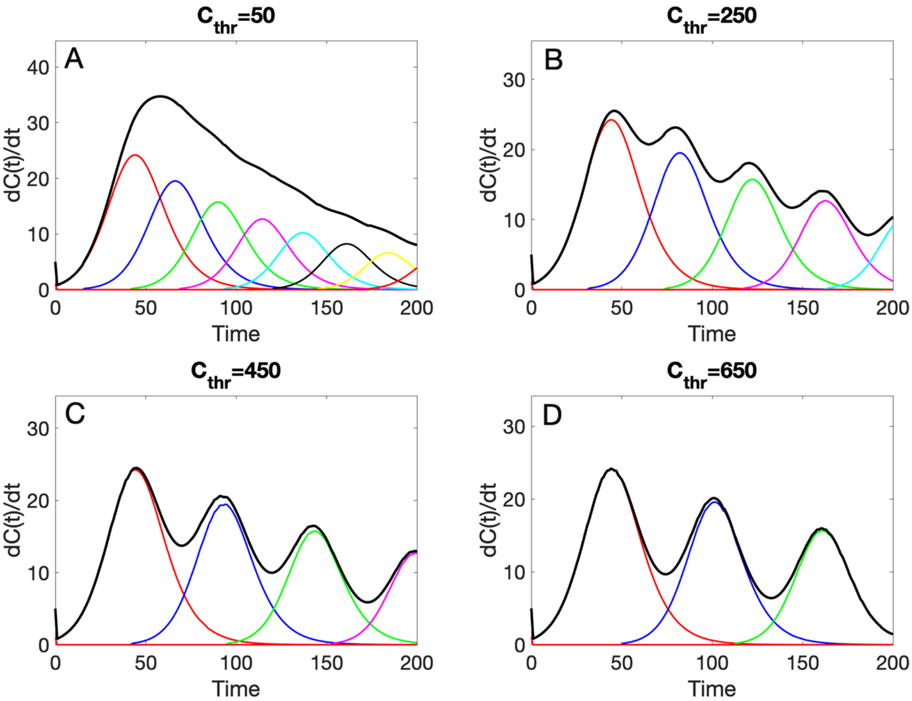


Fig. 3 Four representative profiles of the spatial wave sub-epidemic model where the sub-epidemic building block is modeled using the generalized logistic growth model and characterized by the following parameters: $r = 0.18, p = 0.18, K = 1000, q = 0.24, n = 8$, and the C_{thr} value is varied with values: **A** 50, **B** 250, **C** 450, **D** 650. An exponential function is used to model the decline of sub-epidemic sizes (<`typedecline1`>=1). The solid black line corresponds to the overall aggregated curve whereas the individual sub-epidemics are shown in different colors

(`<flag1>=1`) and the following model parameter values: $r = 0.18, p = 0.18, K = 1000, q = 0.24, n = 8, C_{thr} = 50$. The initial condition $C(0) = 5$, and the total duration of the simulation is set at 200.

```
>> plot_SW_subepidemic(1,0.18,0.9,[],1000,0.24,8,50,1,5,200)
```

Of note, in the above call, the value of parameter a is passed empty ([]) since the generalized logistic growth model does not use this parameter. This function will generate a figure (Fig. 3A) that shows the corresponding model solution $dC(t)/dt$. Additional representative simulations with other values of the C_{thr} are shown in Fig. 3.

In the next section, we describe four comprehensive performance metrics that can be used to assess both calibration and forecasting performance. Specifically, the mean absolute error (MAE) and the mean squared error (MSE) are used to assess the performance of point forecasts, while the coverage of the 95% prediction interval (95% PI) and the weighted interval score (WIS) evaluate the performance of distributional forecasts by accounting for uncertainty in model fit and predictions.

Performance metrics

To assess the performance of the models during the calibration or forecasting periods, we used four performance metrics: the mean absolute error (MAE), the mean squared error (MSE), the coverage of the 95% prediction intervals (95% PI), and the weighted interval score (WIS) [27]. While it is possible to generate h-time units ahead forecasts of an evolving process, those forecasts looking into the future can only be evaluated until sufficient data for the h-time units ahead has been collected. In the `options_forecast.m` file, the parameter `<get-performance>` is a Boolean variable (0/1) to indicate whether the user wishes to compute the performance metrics of the forecasts when sufficient data is available.

The *mean absolute error* (MAE) is given by:

$$\text{MAE} = \frac{1}{N}\sum\nolimits_{h=1}^{N}\left|f\left(t_h, \widehat{\Theta}\right) - y_{t_h}\right|,$$

where $t_h$are the time points of the time series data [28], and N is the calibration or forecasting period length. Similarly, the *mean squared error* (MSE) is given by:

$$\text{MSE} = \frac{1}{N}\sum\nolimits_{h=1}^{N}(f\left(t_h, \widehat{\Theta}\right) - y_{t_h})^2,$$

where $t_h$are the time points of the time series data [28], and N is the calibration or forecasting period length. The coverage of the *95% prediction interval* (PI) corresponds to the fraction of data points that fall within the 95% PI, calculated as

$$\text{95\% PI coverage} = \frac{1}{N}\sum\nolimits_{h=1}^{N} 1\{Y_{t\,h} > L_{t_h} \cap Y_{t_h} < U_{t_h}\}$$

where $L_{t_h}$and U_{t_h} are the lower and upper bounds of the 95% PIs, respectively, $Y_{t_h}$are the data and 1 is an indicator variable that equals 1 if $Y_{t_h}$is in the specified interval and 0 otherwise.

The *weighted interval score* (WIS) [27, 29], which is a proper score recently embraced for quantifying model forecasting performance in epidemic forecasting studies [30–33], provides quantiles of predictive forecast distribution by combining a set of Interval Scores (IS) for probabilistic forecasts. An IS is a simple proper score that requires only a central $(1 - \alpha) \times 100\%$ PI [27] and is described as

$$IS_\alpha\left(F, y\right) = (u - l) + \frac{2}{\alpha} \times \left(l - y\right) \times 1\left(y < l\right) + \frac{2}{\alpha} \times \left(y - u\right) \times 1\left(y > u\right).$$

In this Eq. 1 refers to the indicator function, meaning that $1\left(y < l\right) = 1$ if $y < l$ and 0 otherwise. The terms l and u represent the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the forecast F. The IS consists of three distinct quantities:

1. The sharpness of F, given by the width $u - l$ of the central $(1 - \alpha) \times 100\%$ PI.
2. A penalty term $\frac{2}{\alpha} \times \left(l - y\right) \times 1\left(y < l\right)$ for the observations that fall below the lower end point l of the $(1 - \alpha) \times 100\%$ PI. This penalty term is directly proportional to the distance between y and the lower end l of the PI. The strength of the penalty depends on the level α.
3. An analogous penalty term $\frac{2}{\alpha} \times \left(y - u\right) \times 1\left(y > u\right)$ for the observations falling above the upper limit u of the PI.

To provide more detailed and accurate information on the entire predictive distribution, we report several central PIs at different levels $(1 - \alpha_1) < (1 - \alpha_2) < \cdots < (1 - \alpha_K)$ along with the predictive median, $\underset{\sim}{y}$, which can be seen as a central prediction interval at level $1 - \alpha_0 \rightarrow 0$. This is referred to as the WIS, and it can be evaluated as follows:

$$WIS_{\alpha_{0:K}}\left(F, y\right) = \frac{1}{K + \frac{1}{2}}\cdot\left(w_0.\left|y - \tilde{y}\right| + \sum_{k=1}^{K} w_k.IS_{\alpha_k}\left(F, y\right)\right),$$

where, $w_k = \frac{\alpha_k}{2}$for $k = 1,2,\ldots.K$ and $w_0 = \frac{1}{2}$. Hence, WIS can be interpreted as a measure of how close the

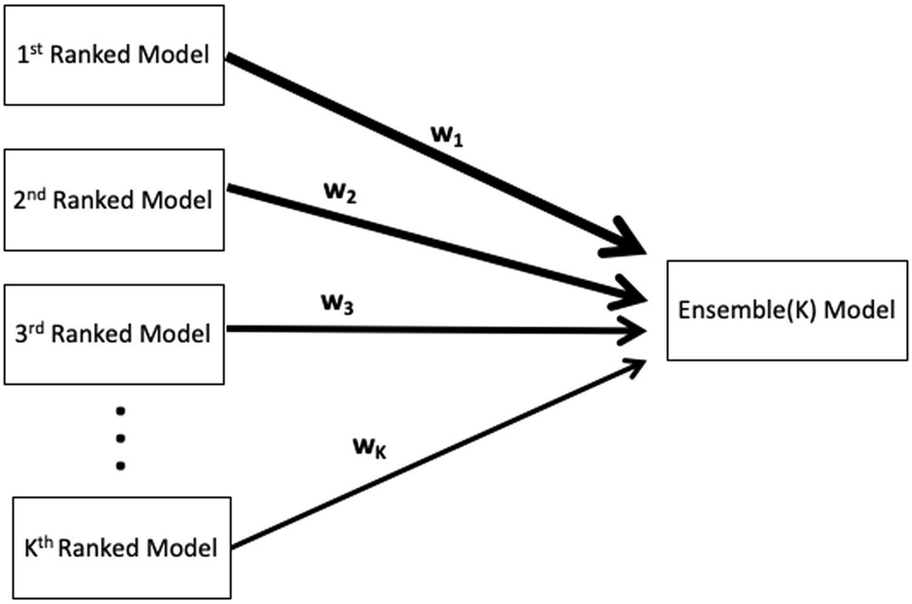


Fig. 4 Schematic diagram of the construction of the ensemble model from the weighted combination of the highest-ranking sub-epidemic models as deemed by the AIC_{c_k} for the *k*-th model where $AIC_{c_1} \leq \cdots \leq AIC_{c_K}$ and *k* = 1, …, *K*. An ensemble derived from the top-ranking *K* models is denoted by Ensemble(K)

entire distribution is to the observation in units on the scale of the observed data [31, 34].

Doubling times

Doubling times characterize the sequence of times at which the cumulative incidence doubles. We denote the times at which cumulative incidence doubles by t_{d_j}, such that $2C(t_{d_j}) = C(t_{d_{j+1}})$ where $t_{d_0} = 0, C\left(t_{d_0}\right) = C_0$, $j = 1,2,3,\ldots,n_g$ and n_g is the total number of times cumulative incidence doubles [35]. The actual sequence of "doubling times" is defined as follows:

$$d_j = \Delta t_{d_j} = t_{d_j} - t_{d_{j-1}} \text{where } j = 1,2,3,\ldots,n_g.$$

For exponential growth, doubling times remain invariant and are given by $(ln2)/r$, whereas the doubling times increase when the growth pattern follows sub-exponential growth [36]. We can characterize the doubling times and their uncertainty from the best-fit model $f\left(t,\widehat{\Theta}\right)$ [37]. We can evaluate the uncertainty of the sequence of doubling times and the overall doubling time using the model parameter estimates derived from bootstrapping $\left(\widehat{\Theta}_b\right)$, where $b = 1,2,3,\ldots,B$. That is, $d_j\left(\widehat{\Theta}_b\right)$ provides a sequence of doubling times for a set of bootstrap parameter estimates, $\widehat{\Theta}_b$, where $b = 1,2,3,\ldots,B$. We can use these curves to derive 95% CIs for the sequence of doubling times and quantify the probability of observing a given number of doublings.

Constructing ensemble forecasts from top-ranking models

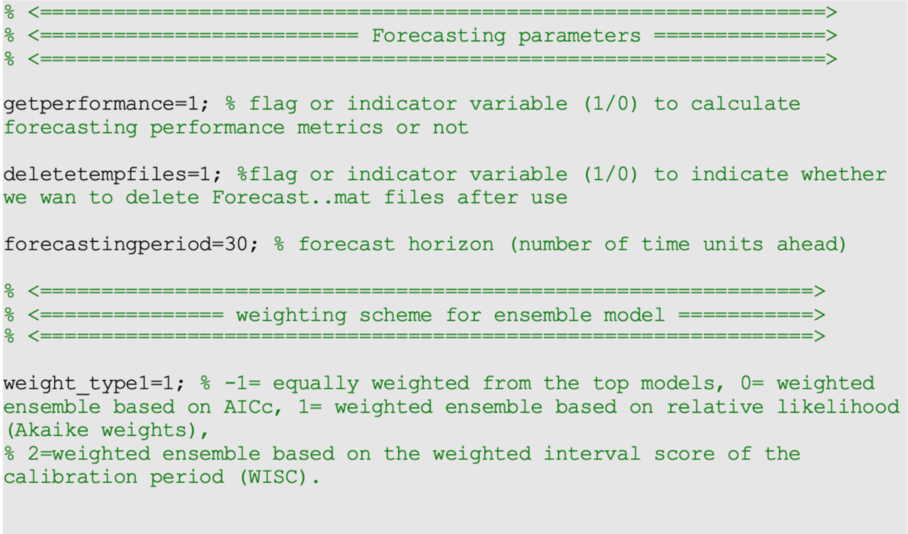
Ensemble models that combine the strength of multiple models may exhibit significantly enhanced predictive performance (e.g [11, 17, 38, 39]). An ensemble model derived from the top-ranking I models is denoted by the Ensemble(1), illustrated in Fig. 4. Thus, Ensemble(2) and Ensemble(3) refer to the ensemble models generated from the combination of the top-ranking 2 and 3 models, respectively. The ensemble models can be derived from the unweighted (equal weights across contributing individual models) or a weighted combination of the highest-ranking sub-epidemic models based on the quality of fit as deemed by the $AIC_{c_i}$for the i-th model where $AIC_{c_1} \leq \cdots \leq AIC_{c_I}$ and i = 1, …, I. In this case, we compute the weight $w_i$for the i-th model, i = 1, …, I, where $\sum w_i = 1$ as follows:

$$w_i = \frac{\frac{1}{AIC_{c_i}}}{\frac{1}{AIC_{c_1}} + \frac{1}{AIC_{c_2}} + \cdots + \frac{1}{AIC_{c_I}}} \text{for all } i = 1,2,\ldots,I,$$

and hence $w_I \leq \ldots \leq w_1$.

The estimated mean curve of daily COVID-19 cases for the Ensemble(I) model is:

$$f_{ens(I)}(t) = \sum_{i=1}^{I} w_i f_i\left(t,\widehat{\Theta}^{(i)}\right),$$



```
% <=================================================================>
% <========================== Forecasting parameters ==============>
% <=================================================================>

getperformance=1; % flag or indicator variable (1/0) to calculate
forecasting performance metrics or not

deletetempfiles=1; %flag or indicator variable (1/0) to indicate whether
we wan to delete Forecast..mat files after use

forecastingperiod=30; % forecast horizon (number of time units ahead)

% <================================================================>
% <=============== weighting scheme for ensemble model ===========>
% <================================================================>

weight_type1=1; % -1= equally weighted from the top models, 0= weighted
ensemble based on AICc, 1= weighted ensemble based on relative likelihood
(Akaike weights),
% 2=weighted ensemble based on the weighted interval score of the
calibration period (WISC).
```

Fig. 5 Contents of `options_forecast.m` file, that specify the parameters related to the epidemic forecasts including the forecasting horizon and the type of ensemble weights (e.g., unweighted, weighted based on AIC_c, weighted based on the relative likelihood of the models, and weighted based on the WIS of the calibration period)

where given the training data, $\widehat{\Theta}^{(i)}$ denotes the set of estimated parameters, and $f_i\left(t,\widehat{\Theta}\right)^{(i)}$ denotes the estimated mean curve of daily COVID-19 cases, for the i-th model. Accordingly, we compute the weighted average and sample the bootstrap realizations of the forecasts for each model to construct the 95% CI or PI using the 2.5% and 97.5% quantiles [17]. Alternatively, we can set the ensemble weights based on different calibration performance metrics for the top-ranked models. For instance, we can make the ensemble weights proportional to the relative likelihood (l) rather than the reciprocal of the AIC_c. Let AIC_{min} denote the minimum AIC from the set of models. The relative likelihood of model i is given by $l_i = e^{((AIC_{min}-AIC_i)/2)}$ [40]. We compute the weight w_i for the i-th model where $\sum w_i = 1$ as follows:

$$w_i = \frac{l_i}{l_1 + l_2 + \cdots + l_I}\text{forall}i = 1,2,\ldots,I,$$

and hence $w_I \leq \ldots \leq w_1$.

In the `options_forecast.m file`, we can specify four types of ensemble weights using `<weight_type1>`. Specifically, unweighted (`<weigth_type1>=-1`), weighted according to the AIC_c (`<weight_type1>=0`), weighted based on the relative likelihood (`weight_type1=1`), weighted based on the reciprocal of the WIS metric of the calibration period (`<weight_type1>=2`).

In the `options_forecast.m` file, we can specify the parameters related to the epidemic forecasts, including the forecasting horizon and the type of ensemble weights (Fig. 5).

Results and discussion

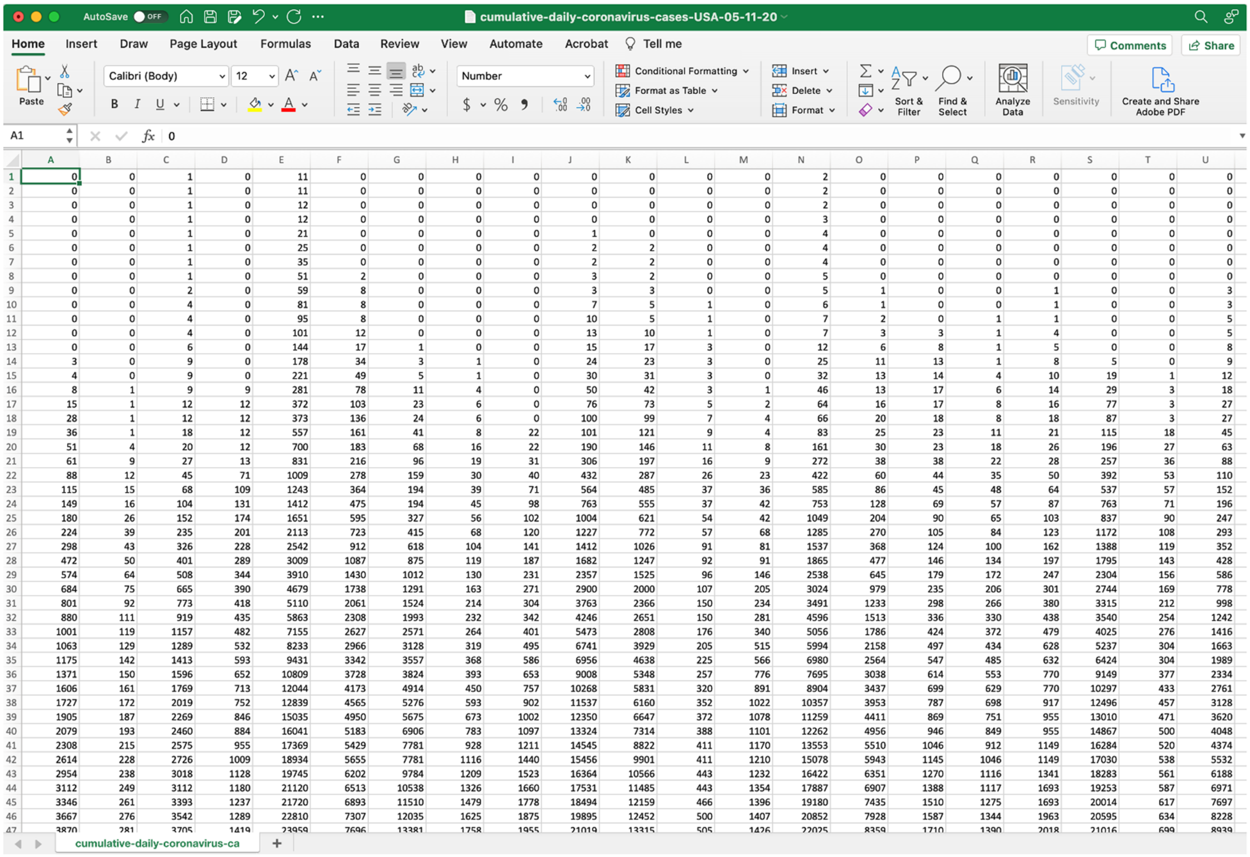
The dataset

The time series data file is a text file with the extension *.txt in the input folder. The data file can contain one or more incidence time series (one per column). Each column corresponds to the incidence curve over time for each epidemic corresponding to a different area/group. For instance, each column could contain time series data corresponding to different U.S. states or countries worldwide. In the `options.m` file, a specific data column can be accessed for inference using the parameter `<outbreakx>`. If the time series file contains cumulative incidence count data, the name of the file containing the time series data starts with "cumulative" according to the following format:

cumulative`-< cadtemporal>-<caddisease>-<datatype>-<cadregion>-<caddate1>`**.txt**.

where `<cadtemporal>` is a string parameter that indicates the temporal resolution of the data (e.g., daily, weekly, yearly). Parameter `<caddisease>` is a string used to indicate the name of the disease related to the time series data, `<datatype>` is a string parameter indicating the nature of the data (e.g., cases, deaths, and hospitalizations), whereas `<cadregion>` is a string parameter indicating the geographic region of the time series contained in the file (New York, USA, World, Asia, Africa). Finally, `<caddate1>` is a string to indicate the date for the most recent observation in the data file with the format: `mm-dd-yyyy`.

To illustrate the methodology presented in this tutorial paper, we used daily COVID-19 cases reported in



A1 f_x 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0	0	1	0	11	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
2	0	0	1	0	11	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
3	0	0	1	0	12	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
4	0	0	1	0	12	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
5	0	0	1	0	21	0	0	0	0	1	0	0	0	4	0	0	0	0	0	0	0
6	0	0	1	0	25	0	0	0	0	2	2	0	0	4	0	0	0	0	0	0	0
7	0	0	1	0	35	0	0	0	0	2	2	0	0	4	0	0	0	0	0	0	0
8	0	0	1	0	51	2	0	0	0	3	2	0	0	5	0	0	0	0	0	0	0
9	0	0	2	0	59	8	0	0	0	3	3	0	0	5	1	0	0	1	0	0	3
10	0	0	4	0	81	8	0	0	0	7	5	1	0	6	1	0	0	1	0	0	3
11	0	0	4	0	95	8	0	0	0	10	5	1	0	7	2	0	1	1	0	0	5
12	0	0	4	0	101	12	0	0	0	13	10	1	0	7	3	3	1	4	0	0	5
13	0	0	6	0	144	17	1	0	0	15	17	3	0	12	6	8	1	5	0	0	8
14	3	0	9	0	178	34	3	1	0	24	23	3	0	25	11	13	1	8	5	0	9
15	4	0	9	0	221	49	5	1	0	30	31	3	0	32	13	14	4	10	19	1	12
16	8	1	9	9	281	78	11	4	0	50	42	3	1	46	13	17	6	14	29	3	18
17	15	1	12	12	372	103	23	6	0	76	73	5	2	64	16	17	8	16	77	3	27
18	28	1	12	12	373	136	24	6	0	100	99	7	4	66	20	18	8	18	87	3	27
19	36	1	18	12	557	161	41	8	22	101	121	9	4	83	25	23	11	21	115	18	45
20	51	4	20	12	700	183	68	16	22	190	146	11	8	161	30	23	18	26	196	27	63
21	61	9	27	13	831	216	96	19	31	306	197	16	9	272	38	38	22	28	257	36	88
22	88	12	45	71	1009	278	159	30	40	432	287	26	23	422	60	44	35	50	392	53	110
23	115	15	68	109	1243	364	194	39	71	564	485	37	36	585	86	45	48	64	537	57	152
24	149	16	104	131	1412	475	194	45	98	763	555	37	42	753	128	69	57	87	763	71	196
25	180	26	152	174	1651	595	327	56	102	1004	621	54	42	1049	204	90	65	103	837	90	247
26	224	39	235	201	2113	723	415	68	120	1227	772	57	68	1285	270	105	84	123	1172	108	293
27	298	43	326	228	2542	912	618	104	141	1412	1026	91	81	1537	368	124	100	162	1388	119	352
28	472	50	401	289	3009	1087	875	119	187	1682	1247	92	91	1865	477	146	134	197	1795	143	428
29	574	64	508	344	3910	1430	1012	130	231	2357	1525	96	146	2538	645	179	172	247	2304	156	586
30	684	75	665	390	4679	1738	1291	163	271	2900	2000	107	205	3024	979	235	206	301	2744	169	778
31	801	92	773	418	5110	2061	1524	214	304	3763	2366	150	234	3491	1233	298	266	380	3315	212	998
32	880	111	919	435	5863	2308	1993	232	342	4246	2651	150	281	4596	1513	336	330	438	3540	254	1242
33	1001	119	1157	482	7155	2627	2571	264	401	5473	2808	176	340	5056	1786	424	372	479	4025	276	1416
34	1063	129	1289	532	8233	2966	3128	319	495	6741	3929	205	515	5994	2158	497	434	628	5237	304	1663
35	1175	142	1413	593	9431	3342	3557	368	586	6956	4638	225	566	6980	2564	547	485	632	6424	304	1989
36	1371	150	1596	652	10809	3728	3824	393	653	9008	5348	257	776	7695	3038	614	553	770	9149	377	2334
37	1606	161	1769	713	12044	4173	4914	450	757	10268	5831	320	891	8904	3437	699	629	770	10297	433	2761
38	1727	172	2019	752	12839	4565	5276	593	902	11537	6160	352	1022	10357	3953	787	698	917	12496	457	3128
39	1905	187	2269	846	15035	4950	5675	673	1002	12350	6647	372	1078	11259	4411	869	751	955	13010	471	3620
40	2079	193	2460	884	16041	5183	6906	783	1097	13324	7314	388	1101	12262	4956	946	849	955	14867	500	4048
41	2308	215	2575	955	17369	5429	7781	928	1211	14545	8822	411	1170	13553	5510	1046	912	1149	16284	520	4374
42	2614	228	2726	1009	18934	5655	7781	1116	1440	15456	9901	411	1210	15078	5943	1145	1046	1149	17030	538	5532
43	2954	238	3018	1128	19745	6202	9784	1209	1523	16364	10566	443	1232	16422	6351	1270	1116	1341	18283	561	6188
44	3112	249	3112	1180	21120	6513	10538	1326	1660	17531	11485	443	1354	17887	6907	1388	1117	1693	19253	587	6971
45	3346	261	3393	1237	21720	6893	11510	1479	1778	18494	12159	466	1396	19180	7435	1510	1275	1693	20014	617	7697
46	3667	276	3542	1289	22810	7307	12035	1625	1875	19895	12452	500	1407	20852	7928	1587	1344	1963	20595	634	8228
47	3870	281	3705	1419	23959	7696	13381	1758	1955	21019	13315	505	1426	22025	8359	1710	1390	2018	21016	699	8939

Fig. 6 Example data file named `cumulative-daily-coronavirus-cases-USA-05-11-2020.txt` located in the input folder. A partial view in Excel of the contents of the data file is shown

the USA from the publicly available data tracking system of the Johns Hopkins Center for Systems Science and Engineering (CSSE) [41]. The data is also publicly available in the GitHub repository [42]. An example of a data file that we will use in this tutorial is provided in Fig. 6.

If the time series file contains incidence data, the name of the data file does not start with the word 'cumulative' and follows the format:

`<cadtemporal>-<caddisease>-<datatype>-<cadregion>-<caddate1>``.`**`txt`**

For example: `daily-coronavirus-cases-USA-05-11-2020.txt`

In the `options.m` file, the parameter `<datevecfirst1>` is a 3-value vector that specifies the date corresponding to the first data point in time series data in the format: `[yyyy mm dd]`. Similarly, the parameter `<datevecend1>` is a 3-value vector that specifies the date of the most recent data file in the format: `[yyyy mm dd]`. The file

`cumulative-<cadtemporal>-<caddisease>-<datatype>-<cadregion>-`**`<datevecend1>`**`.txt`

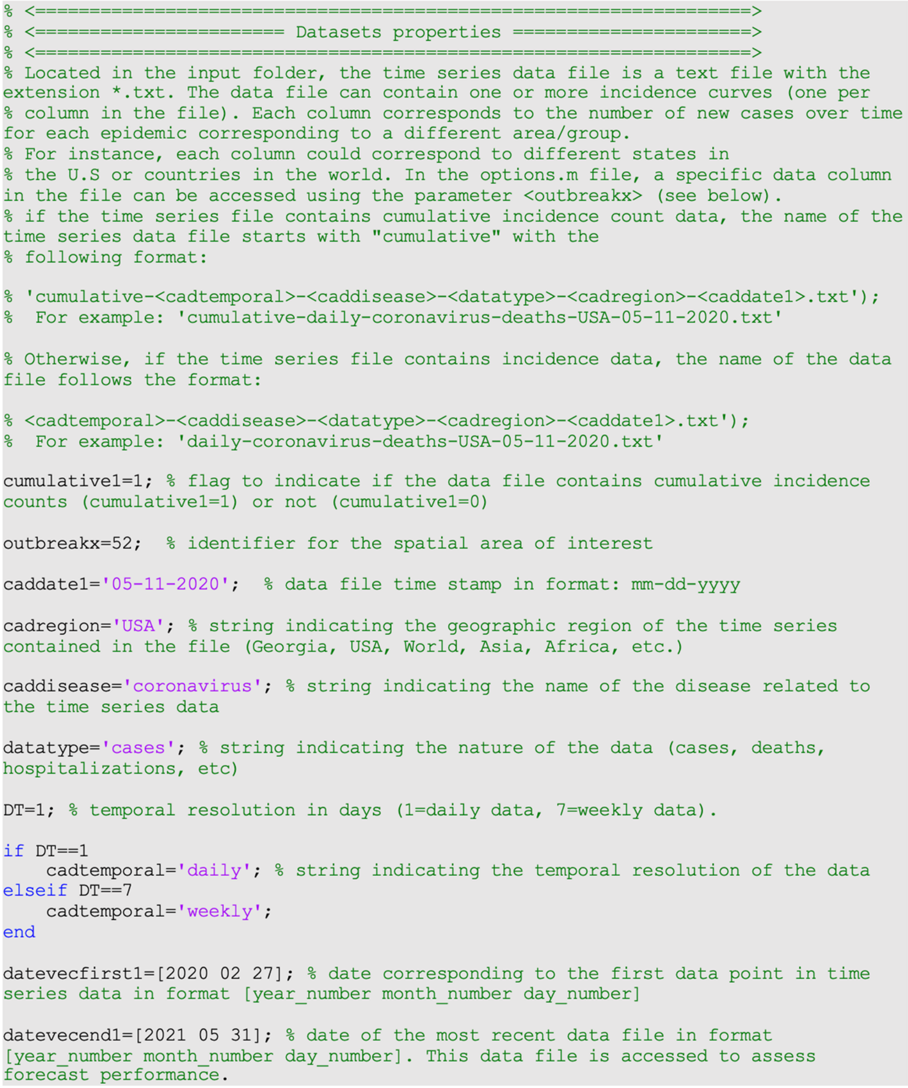
in the input folder with the date **`<datevecend1>`** contains the most recent time series data and is needed to assess forecast performance. Finally, the parameter `<DT>` is an integer indicating the temporal resolution of the time series data (e.g., `<DT>`=1 for daily data; `<DT>`=7 for weekly data) (Fig. 7).

Data adjustments

Data smoothing

To reduce the noise in the original data due to artificial reasons such as the weekend effects, we can smooth out the time series data using the moving average of the time series whereby `<smoothfactor1>` is a parameter in the `options.m` file that specifies the span of the moving average (e.g., `<smoothfactor1>`=1 implies no smoothing applied to the data). Let

$$y_{t_j}=y_{t_1},y_{t_2},\ldots,y_{t_{n_d}} \text{ where } j=1,2,\ldots,n_d$$



```
% <=============================================================================>
% <========================= Datasets properties ======================>
% <=============================================================================>
% Located in the input folder, the time series data file is a text file with the
extension *.txt. The data file can contain one or more incidence curves (one per
% column in the file). Each column corresponds to the number of new cases over time
for each epidemic corresponding to a different area/group.
% For instance, each column could correspond to different states in
% the U.S or countries in the world. In the options.m file, a specific data column
in the file can be accessed using the parameter <outbreakx> (see below).
% if the time series file contains cumulative incidence count data, the name of the
time series data file starts with "cumulative" with the
% following format:

% 'cumulative-<cadtemporal>-<caddisease>-<datatype>-<cadregion>-<caddate1>.txt');
%  For example: 'cumulative-daily-coronavirus-deaths-USA-05-11-2020.txt'

% Otherwise, if the time series file contains incidence data, the name of the data
file follows the format:

% <cadtemporal>-<caddisease>-<datatype>-<cadregion>-<caddate1>.txt');
%  For example: 'daily-coronavirus-deaths-USA-05-11-2020.txt'

cumulative1=1; % flag to indicate if the data file contains cumulative incidence
counts (cumulative1=1) or not (cumulative1=0)

outbreakx=52;  % identifier for the spatial area of interest

caddate1='05-11-2020';  % data file time stamp in format: mm-dd-yyyy

cadregion='USA'; % string indicating the geographic region of the time series
contained in the file (Georgia, USA, World, Asia, Africa, etc.)

caddisease='coronavirus'; % string indicating the name of the disease related to
the time series data

datatype='cases'; % string indicating the nature of the data (cases, deaths,
hospitalizations, etc)

DT=1; % temporal resolution in days (1=daily data, 7=weekly data).

if DT==1
    cadtemporal='daily'; % string indicating the temporal resolution of the data
elseif DT==7
    cadtemporal='weekly';
end

datevecfirst1=[2020 02 27]; % date corresponding to the first data point in time
series data in format [year_number month_number day_number]

datevecend1=[2021 05 31]; % date of the most recent data file in format
[year_number month_number day_number]. This data file is accessed to assess
forecast performance.
```

Fig. 7 Contents of `options.m` file, and the values of the parameters related to the data including the temporal resolution of the time series data

denote the smoothed time series of the epidemic trajectory based on the moving average. Here, $t_j$are the time points for the time series data, n_d is the number of observations, and each y_{t_j},

$j = 1, 2, \ldots, n_d$, correspond to the smoothed time series. We recommend that the user set the average to multiples of seven to reduce the weekend effects in the reported data.

For the daily COVID-19 case data employed for illustration purposes, we set<smoothfactor1>=7 and smooth out the daily series using a 7-day moving average to reduce the noise in the original data due to artificial reasons such as the weekend effects.

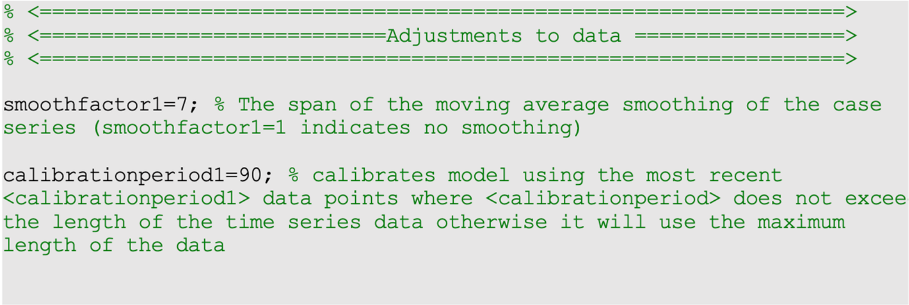
Calibration period

To fit the models to the most recent observations in a time series file, we can specify the length of the calibration period whereby<calibrationperiod1>indicates the number of recent data points that will be used to calibrate the models. If<calibrationperiod1>exceeds the length of the time series in the data file, the calibration period is set to the maximum length of the available data.

For illustration purposes, we used a 90-day calibration period (i.e., <calibrationperiod1> = 90) (Fig. 8).

Fitting the sub-epidemic wave models to data with quantified uncertainty

To fit the sub-epidemic wave models to the data with quantified uncertainty, we need to run the function



```
% <=================================================================>
% <============================Adjustments to data =================>
% <=================================================================>

smoothfactor1=7; % The span of the moving average smoothing of the case
series (smoothfactor1=1 indicates no smoothing)

calibrationperiod1=90; % calibrates model using the most recent
<calibrationperiod1> data points where <calibrationperiod> does not excee
the length of the time series data otherwise it will use the maximum
length of the data
```

Fig. 8 Contents of options.m file, the values of the parameters related to smoothing and calibration period

Run_SW_subepidemicFramework.m. This function uses the input parameters provided by the user in the options.m file. However, the function can also receive <outbreakx> and <caddate1> as passing input parameters while the remaining inputs are obtained from the options.m file.

For example, to fit the ensemble sub-epidemic models to the daily curve of COVID-19 cases in the USA as of the week of '05-11-2020' (data file path: input/cumulative-daily-coronavirus-cases-USA-05-11-2020.txt), we can run the function from MATLAB's command line window as follows:

```
>> Run_SW_subepidemicFramework(52,'05-11-2020')
```

This function will generate several output MATLAB files in the output folder. For instance, the following output file contains the fits of the top-ranking models:

ABC-original-npatchesfixed-4-onsetfixed-0-typedecline-2-smoothing-1-daily-coronavirus-cases-USA-state-52-05-11-2020-flag1-1-method-0-dist-0-calibrationperiod-90.mat

Please note that the names of the output files contain the values of the parameters for reference.

The following output files contain the uncertainty characteristics associated with each of the top-ranking models:

a) modifiedLogisticPatch-original-npatchesfixed-4-onsetfixed-0-typedecline-2-smoothing-1-daily-coronavirus-cases-USA-state-52-05-11-2020-flag1-1-method-0-dist-0-calibrationperiod-90-**rank-1**.mat
b) modifiedLogisticPatch-original-npatchesfixed-4-onsetfixed-0-typedecline-2-smoothing-1-daily-coronavirus-cases-USA-state-52-05-11-2020-flag1-1-method-0-dist-0-calibrationperiod-90-**rank-2**.mat
c) modifiedLogisticPatch-original-npatchesfixed-4-onsetfixed-0-typedecline-2-smoothing-1-daily-coronavirus-cases-USA-state-52-05-11-2020-flag1-1-method-0-dist-0-calibrationperiod-90-**rank-3**.mat
d) modifiedLogisticPatch-original-npatchesfixed-4-onsetfixed-0-typedecline-2-smoothing-1-daily-coronavirus-cases-USA-state-52-05-11-2020-flag1-1-method-0-dist-0-calibrationperiod-90-**rank-4**.mat

These output internal files are needed to plot model fits, derive parameter estimates, generate short-term forecasts, and quantify the calibration and forecasting performance metrics.

Plot the mean model fits and quality of fit metrics for the top-ranked models

After running the function Run_SW_subepidemicFramework.m with the desired input parameters, we can use the function plotRankings_SW_subepidemicFramework.m to plot the mean model fits of the top-ranking models including their sub-epidemic profiles and the associated quality of model fit metrics including the AIC_c, the relative likelihood, and the evidence ratio based on the inputs. However, this function can also receive <outbreakx> and <caddate1> as passing input parameters while the remaining inputs are obtained from the options.m file. Running this function from MATLAB's command line, we have:

```
>>plotRankings_SW_subepidemicFramework(52,'05-11-2020')
```

Figures 9 and 10 illustrate the outputs obtained from this function call. Figure 9 shows the mean model fits of the top-ranked sub-epidemic models, which indicates that the 1st-ranked model consists of 3 sub-epidemics. In contrast, the 2nd, 3rd, and 4th -ranked sub-epidemic

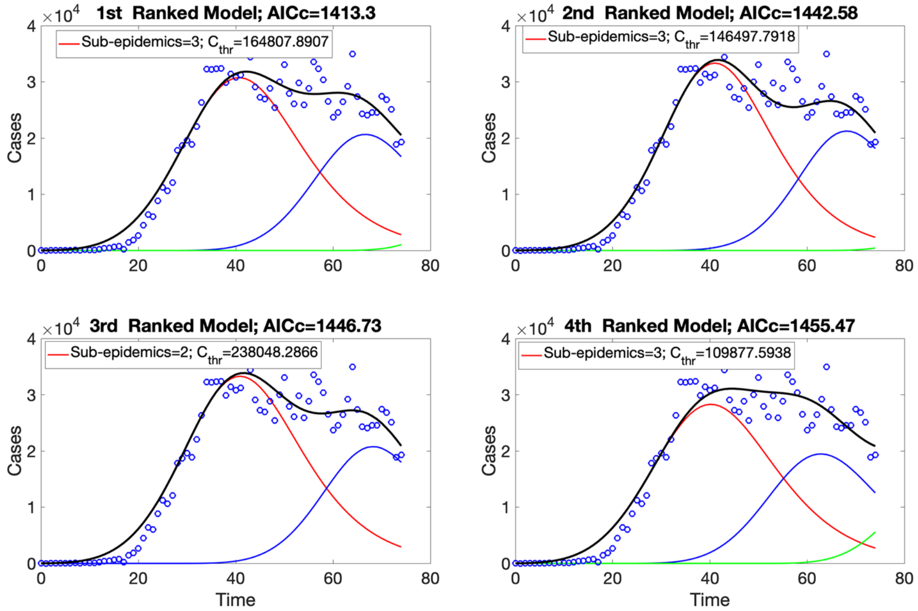


Fig. 9 Mean model fits of the top-ranked sub-epidemic models (<`topmodelsx`>=`4` in `options.m` file) calibrated to the daily curve of COVID-19 cases in the USA from 27-Feb-2020 to 11-May-2020. The solid lines of blue, red, and green correspond to the individual sub-epidemic curves. The solid black line represents the overall aggregated epidemic curve. The legend in each panel indicates the number of sub-epidemics involved in each model and the value of the C_{thr}parameter

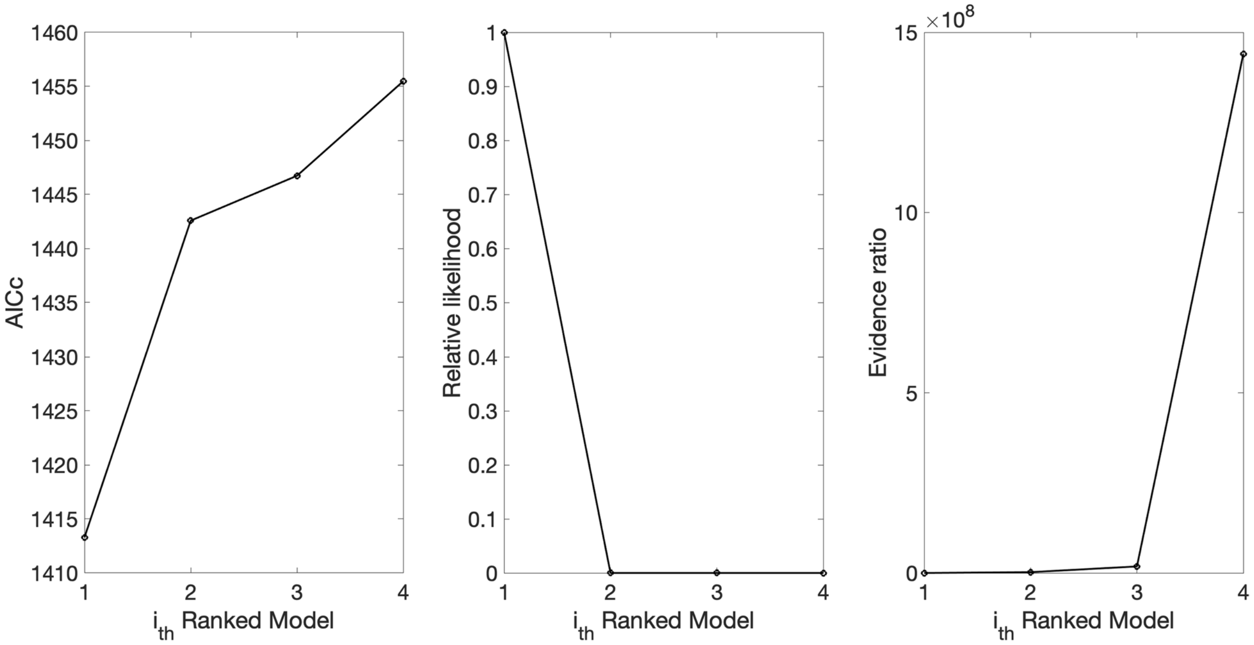


Fig. 10 Quality of model fit metrics for the top-ranked sub-epidemic models (<`topmodelsx`>=`4` in `options.m` file) calibrated to the daily curve of COVID-19 cases in the USA from 27-Feb-2020 to 11-May-2020

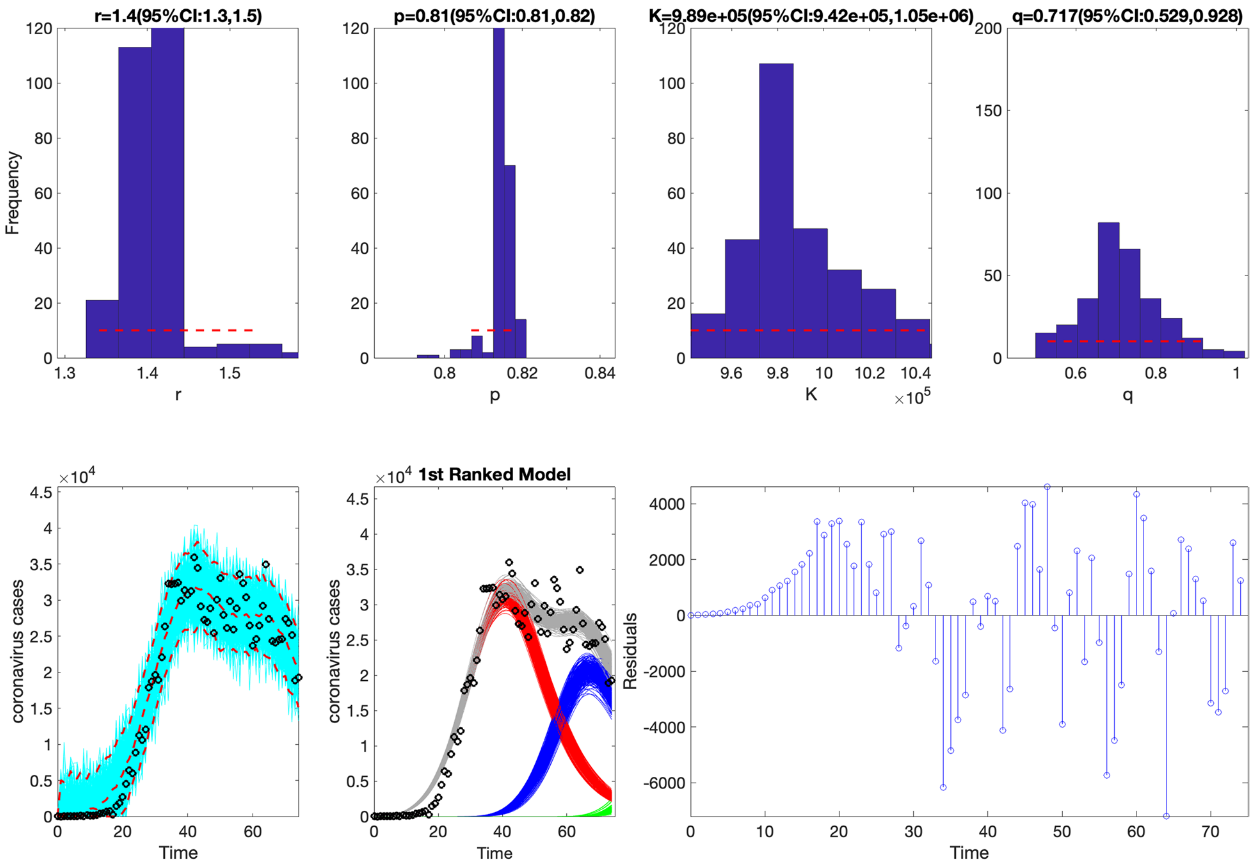


Fig. 11 Fit of the 1st-ranked sub-epidemic wave model to the daily curve of COVID-19 cases in the USA from 27-Feb-2020 to 11-May-2020. The model captures the entire epidemic period well, including the broad peak dynamics, by integrating three asynchronous sub-epidemics. The best model fit (solid red line) and 95% prediction interval (dashed red lines) are shown. The cyan curves correspond to the associated uncertainty from individual bootstrapped curves, which are used to derive the 95% prediction intervals. The sub-epidemic mean profiles obtained from the parametric bootstrapping with 300 bootstrap realizations are shown in the center panels. The red, blue, and green curves represent the three sub-epidemic profiles, and the grey curves are the estimated aggregate epidemic trajectories. Black circles correspond to the data points. The empirical distributions of the parameters and the corresponding estimates are shown in the top panels. The residuals are also shown

models consist of 2 sub-epidemics. It is important to note that there was severe underreporting of cases during the early phase of the epidemic. The corresponding goodness of fit statistics of the top-ranked models, including the AIC_c, the relative likelihood, and the evidence ratio, are shown in Fig. 10. It also saves the AIC_c values of the top-ranked models in the following .csv file:

AICc-topRanked-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv.

For comparison, a simpler growth model consisting of a single sub-epidemic (<npatches_fixed>=1) performs substantially worse ($AIC_c = 1530.4$; Supplementary Fig. 1).

Plot the model fits, parameter estimates, and performance metrics of the top-ranking models

Using the function plotFit_SW_subepidemicFramework.m, we can plot the fits of the top-ranking models, including their sub-epidemic profiles, parameter estimates, and residual plots based on the inputs indicated in the options.m file. However, this function can also receive <outbreakx> and <caddate1> as passing input parameters while the remaining inputs are obtained from the options.m file.

In addition, this function also plots the empirical distributions of the parameters associated with each of the top-ranking models and the calibration performance metrics (MSE, MAE, 95% P.I., and WIS). Finally, this

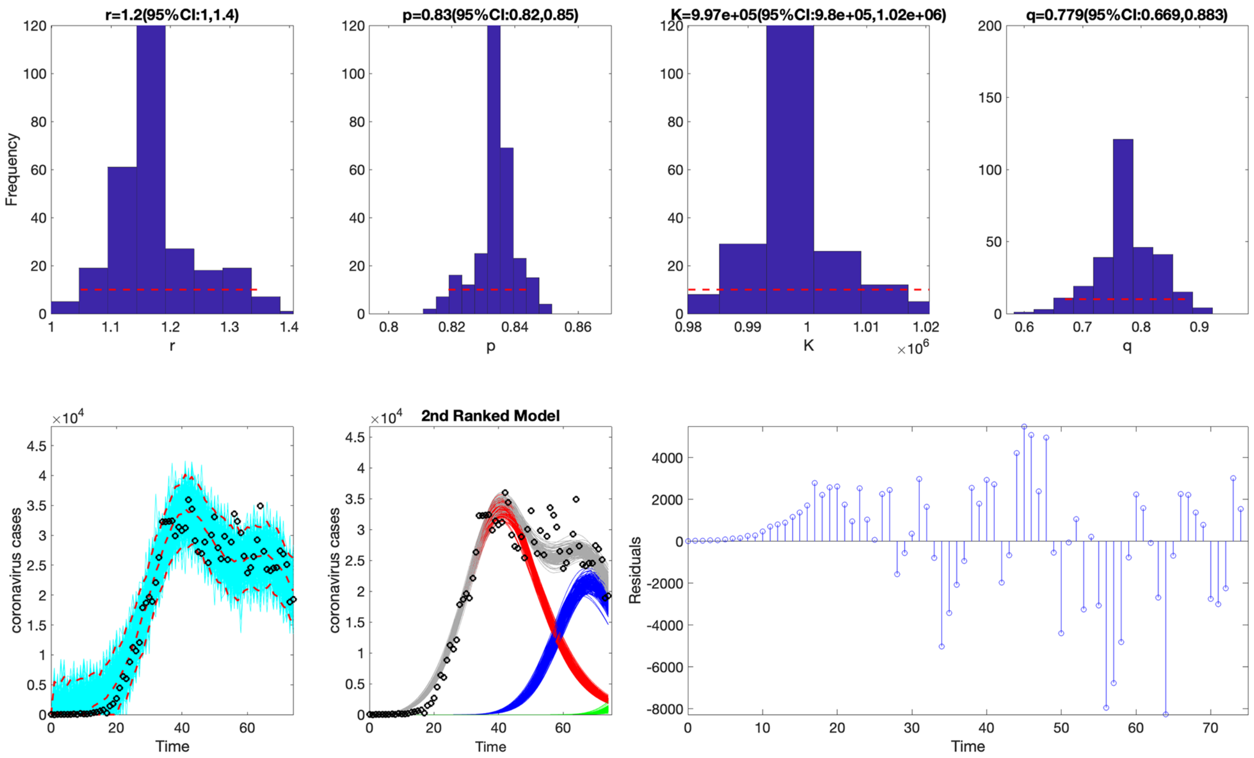


Fig. 12 Fit of the 2nd-ranked sub-epidemic wave model to the daily curve of COVID-19 cases in the USA from 27-Feb-2020 to 11-May-2020. The model captures the entire epidemic period well, including the broad peak dynamics, by integrating three asynchronous sub-epidemics. The best model fit (solid red line) and 95% prediction interval (dashed red lines) are shown. The cyan curves correspond to the associated uncertainty from individual bootstrapped curves, which are used to derive the 95% prediction intervals. The sub-epidemic mean profiles obtained from the parametric bootstrapping with 300 bootstrap realizations are shown in the center panels. The red, blue, and green curves represent the two sub-epidemic profiles, and the grey curves are the estimated aggregate epidemic trajectories. Black circles correspond to the data points. The empirical distributions of the parameters and the corresponding estimates are shown in the top panels. The residuals are also shown

function also outputs `.csv` files in the output folder with the calibration performance metrics, the parameter estimates associated with the top-ranking models, the corresponding Monte Carlo standard errors of the parameters, and the estimated sequence of doubling times for each of the top-ranked models. Using the default parameter values indicated in the `options.m` file, the actual call to this function from MATLAB's command line follows:

```
>> plotFit_SW_subepidemicFramework
```

Figures 11 and 12 illustrate the outputs from the above call to the function. The fits of the 1st and 2nd ranked sub-epidemic models, including the sub-epidemic profiles and residuals, to the daily curve of COVID-19 cases are shown in Figs. 11 and 12. These models yield a similarly good fit to the data. The figures also include the empirical distribution of the parameter estimates. These parameter estimates are well identified as the confidence intervals lie in a well-defined range of values [13]. The calibration performance metrics capturing the quality of fit of the top-ranked sub-epidemic models are also displayed in Fig. 13. For instance, this figure indicates that the coverage of the 95% PIs varied little between ~93% and 95% for the top-ranked models. This function will store the following `.csv` files in the output folder:

1) Model parameter estimates:

```
parameters-topRanked-onsetfixed-
0-typedecline-2-flag1-1-method-
0-dist-0-daily-coronavirus-cases-
USA-area-52-05-11-2020.csv
```

2) Monte Carlo standard errors:

```
MCSES-topRanked-onsetfixed-0-type-
decline-2-flag1-1-method-0-dist-
0-daily-coronavirus-cases-USA-
area-52-05-11-2020.csv
```

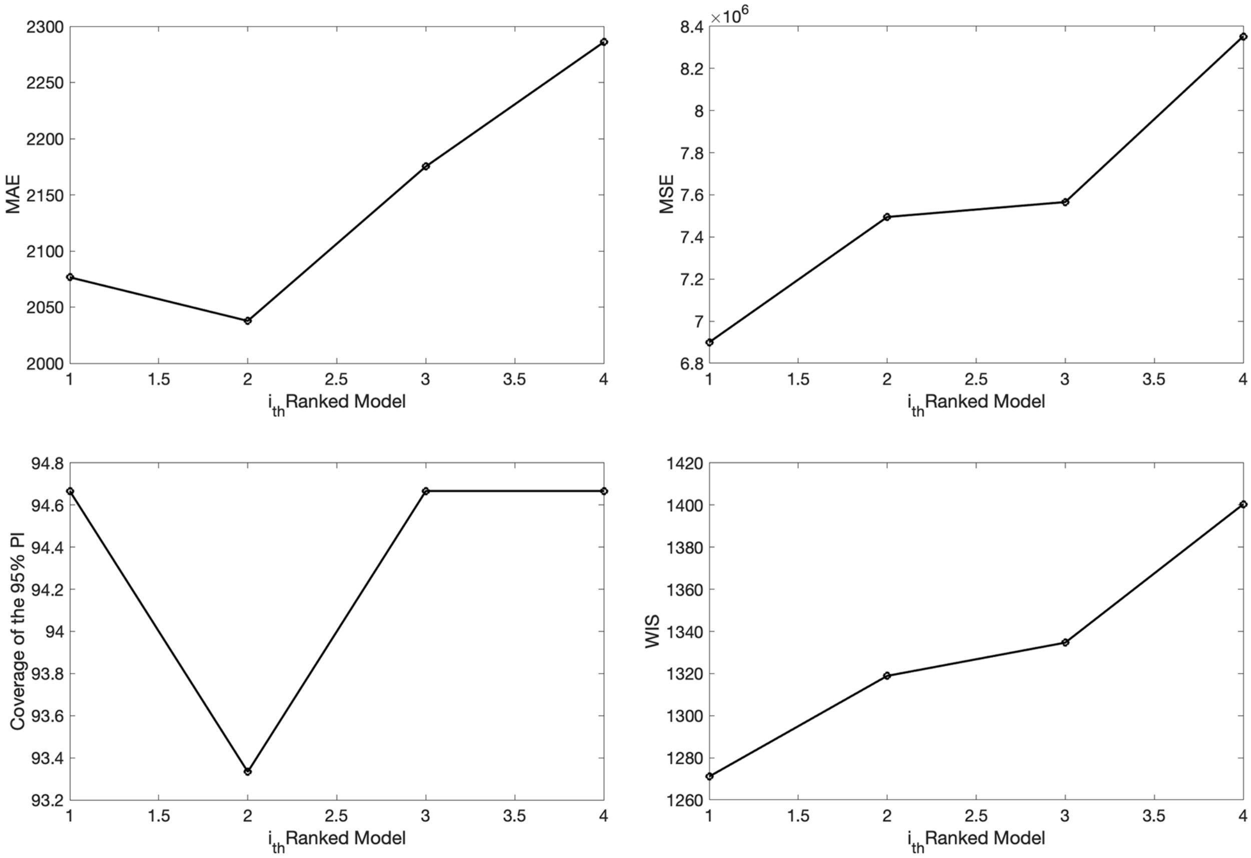


Fig. 13 Calibration performance metrics for the top-ranking sub-epidemic wave models fit to the daily curve of COVID-19 cases in the USA from 27-Feb-2020 to 11-May-2020. These metrics are also saved in a .csv data file (`'performance-calibration-topRanked-onsetfixed-0-typedecline-3-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv'`). For instance, these WIS metrics during the calibration period ranged from ~119.7 to ~124.8 across the four top-ranked models

3) Calibration performance metrics:

`performance-calibration-topRanked-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv`

4) Doubling times for each of the top-ranked models:

doublingTimes-**ranked(1)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-**ranked(2)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-**ranked(3)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-**ranked(4)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

A relevant issue to investigate when using any mathematical model is that of structural or practical parameter identifiability [43]. Structural identifiability arises when one or more model parameters cannot be uniquely estimated using the model, even when the data is free of noise. That is, a lack of structural identifiability is due to issues in the model structure, such as the presence of parameter correlations [12]. On the other hand, practical identifiability occurs when one or more parameters cannot be reliably estimated using the available observed data, which could be associated with the number of

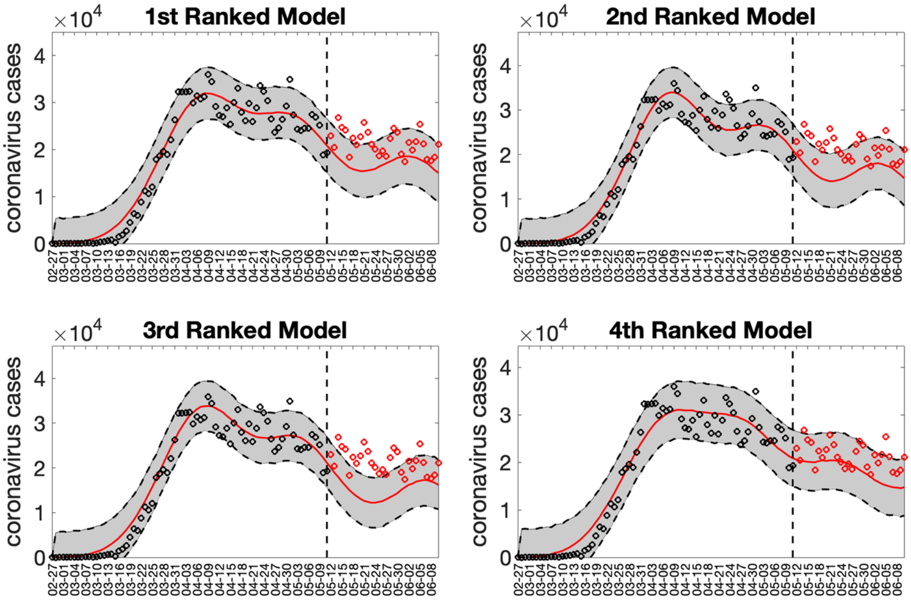


Fig. 14 30-day forecasts derived from the top-ranking sub-epidemic models fit to the daily curve of COVID-19 cases in the USA from 11-May-2020 to 10-June-2020. The model fit (solid line) and 95% prediction interval (shaded area) are also shown. The vertical line indicates the start time of the forecast and separates the calibration and forecast periods. Circles correspond to the data points. Of note, the data associated with each top-ranked model forecast are also saved as .csv files in the output folder

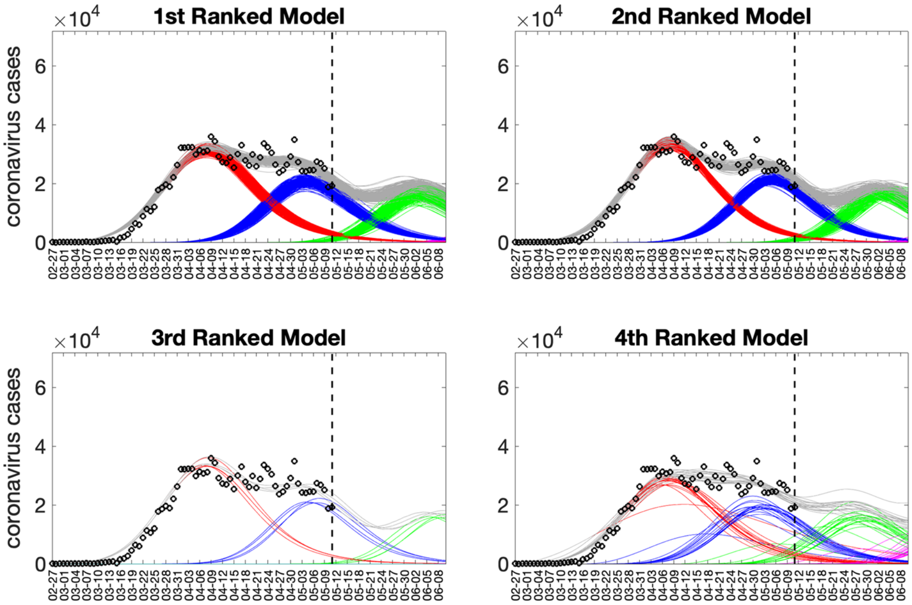


Fig. 15 Sub-epidemic profiles of the 30-day forecasts derived from the top-ranking sub-epidemic models fit to the daily curve of COVID-19 cases in the USA from 11-May-2020 to 10-June-2020. The epidemic wave's sub-epidemic mean curves obtained from the parametric bootstrapping with 300 bootstrap realizations are shown in red, blue, green, and magenta. The gray curves correspond to the overall epidemic trajectory obtained by aggregating the individual sub-epidemic curves. The vertical line indicates the start time of the forecast and separates the calibration and forecast periods

observations available for model calibration and the spatial-temporal resolution of the data. Because the time series of incident cases in the observed epidemic wave is an aggregation of overlapping sub-epidemics, there could be instances when different sub-epidemic profiles may give rise to indistinguishable aggregated epidemic waves as noted elsewhere [44].

Generate the top-ranked and ensemble sub-epidemic model forecasts and the associated forecasting performance metrics

Using the function `plotForecast_SW_subepidemicFramework.m`, we can plot the short-term forecasts from the top-ranking sub-epidemic models and the ensemble models derived from the top-ranking sub-epidemic models based on the inputs indicated in the `options.m` and the `options_forecast.m` files. However, this function can also receive parameters < `outbreakx`>, <`caddate1`>, or <`forecasting period`> as passing input parameters while the remaining inputs are read from the `options.m` and options_forecast.m files. Moreover, the data associated with each top-ranked model and ensemble forecasts are saved as `.csv` files in the output folder.

In addition, this function also plots the forecasting performance metrics (MSE, MAE, 95% P.I., WIS) for the top-ranking models and the ensemble sub-epidemic wave models. Finally, this function also stores *`.csv` files in the output folder with the forecasting performance metrics associated with the top-ranking and ensemble models, and the estimated doubling times for each of the top-ranked models. Using the default parameter values indicated in the `options.m`, and `options_forecast.m` files, the call to this function from MATLAB's command line follows:

```
>>plotForecast_subepidemicFramework
```

Figures 14 and 15 illustrate the outputs obtained from this function call. Figure 14 shows the 30-day forecasts derived from the top-ranking sub-epidemic models, whereas Fig. 15 shows the sub-epidemic profiles of the forecasts. These forecasts indicate that the 1st-ranked model outperformed the other top-ranked models. Moreover, the data associated with the top-ranked model forecasts are also saved as `.csv` files in the output folder.

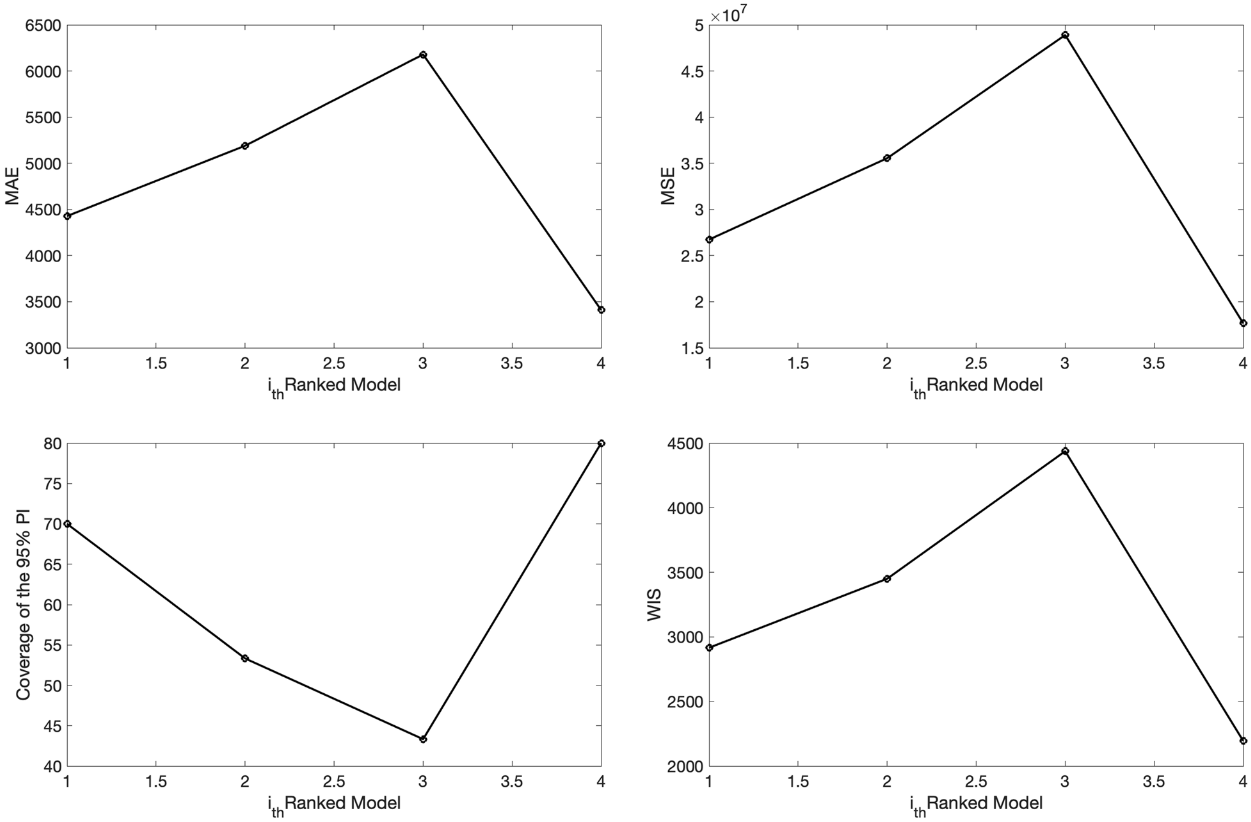


Fig. 16 30-day forecasting performance metrics derived from the top-ranking sub-epidemic models for the daily curve of COVID-19 cases in the USA from 11-May-2020 to 11-June-2020. The forecasting performance metrics are also saved in a .csv data file in the output folder ('performance-forecasting-topRanked-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv')

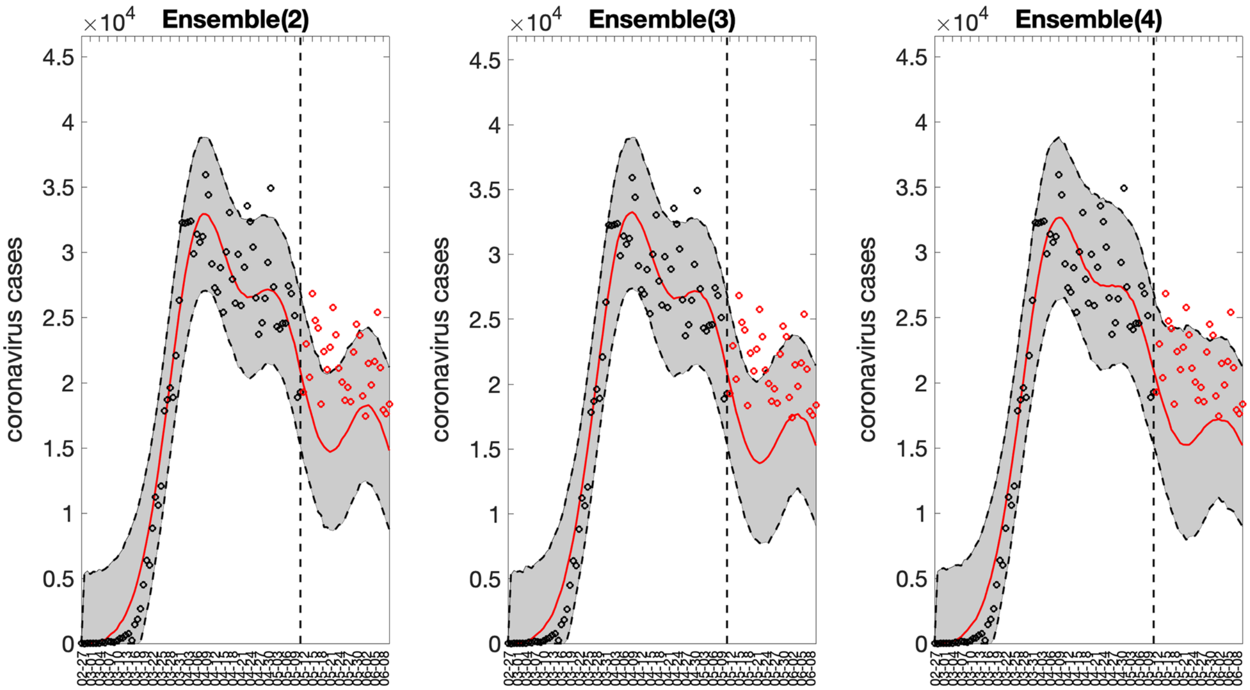


Fig. 17 30-day sub-epidemic ensemble model forecasts (Ensemble(2), Ensemble(3), Ensemble(4)) of COVID-19 cases in the USA from 11-May-2020 to 11-June-2020. Circles correspond to the data points. The model fits (solid line), and 95% prediction intervals (shaded area) are shown. The vertical line indicates the start time of the forecast. Of note, the data associated with each ensemble model forecast are also saved as .csv files in the output folder

The forecasting performance metrics for the top-ranked models are displayed in Fig. 16, and these metrics are also saved in a `.csv` file in the output folder. In comparison, the forecast derived from the simpler growth model consisting of a single sub-epidemic (<`npatches_fixed>=1`) was substantially worse, as shown in Supplementary Fig. 2.

The corresponding 3 ensemble forecasts (Ensemble(2), Ensemble(3), and Ensemble(4)) derived from the weighted combination of the top-ranked models based on their relative likelihood or Akaike weights (`e.g., <weight_type1>=1` in the `options_forecast.m` file) are shown in Fig. 17. Also, the corresponding forecasting performance metrics for the ensemble models are shown in Fig. 18 and are saved in a .csv file in the output folder. The Ensemble(4) performed slightly better than the Ensemble(2) and Ensemble(3) models in terms of the WIS and coverage of the 95% prediction interval. This function will store the following `.csv` files in the output folder:

1) Forecasting performance metrics of the top-ranked models:

```
performance-forecasting-topRanked-
onsetfixed-0-typedecline-2-flag1-
1-method-0-dist-0-horizon-
30-daily-coronavirus-cases-USA-
area-52-05-11-2020.csv
```

2) Forecasting performance metrics of the ensemble models:

```
performance-forecasting-Ensem-
ble-onsetfixed-0-typedecline-
2-flag1-1-method-0-dist-0-horizon-
30-weight_type-1-daily-coronavirus-
cases-USA-area-52-05-11-2020.csv
```

3) Forecasts of the top-ranked models:

```
ranked(1)-onsetfixed-0-typedecline-
2-flag1-1-method-0-dist-0-horizon-
30-daily-coronavirus-cases-USA-
area-52-05-11-2020.csv
ranked(2)-onsetfixed-0-typedecline-
2-flag1-1-method-0-dist-0-horizon-
```

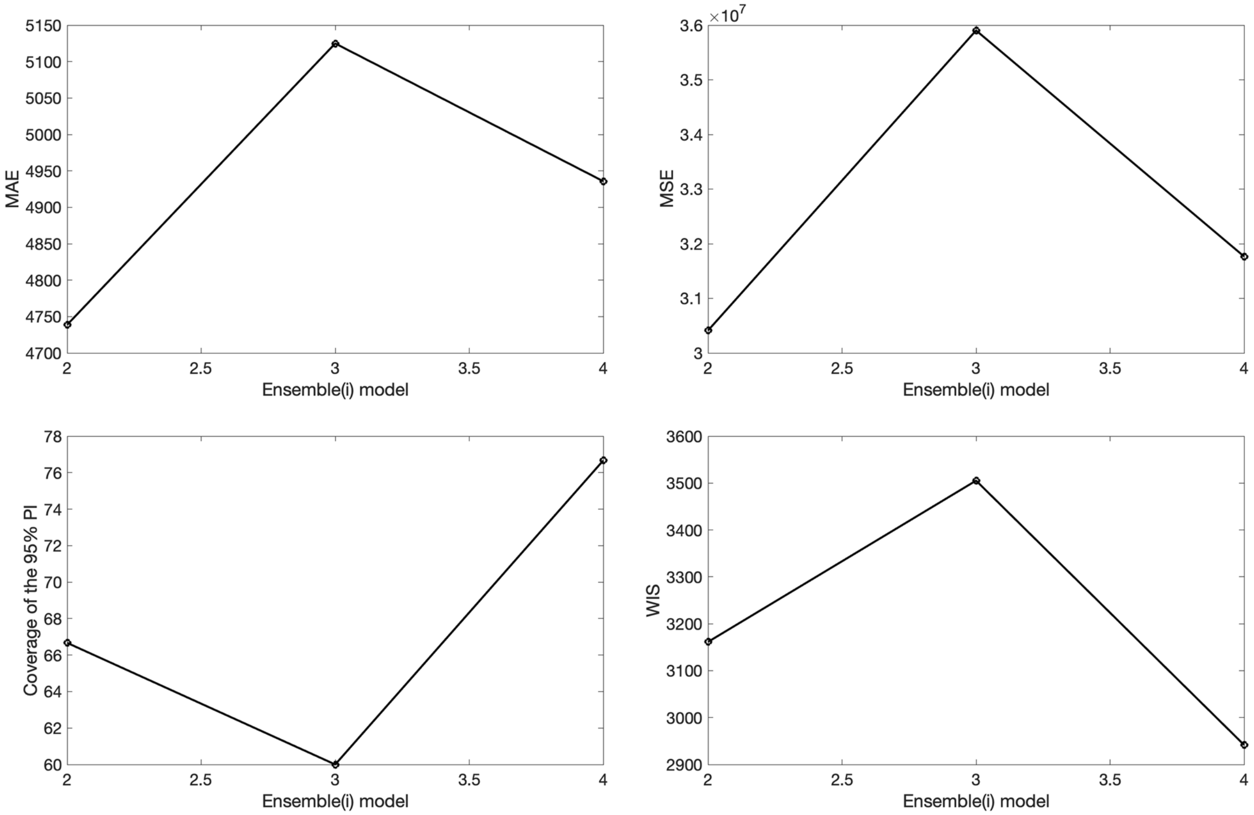


Fig. 18 30-day forecasting performance metrics derived from the sub-epidemic ensemble models for the daily curve of COVID-19 cases in the USA from 11-May-2020 to 11-June-2020. The performance metrics are also saved in a .csv data file in the output folder ('performance-forecasting-Ensemble-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv')

30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

ranked(3)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

ranked(4)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

4) Forecasts of the ensemble models:

Ensemble(2)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

Ensemble(3)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

Ensemble(4)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

5) Sequence of doubling times of the top-ranked models:

doublingTimes-**ranked(1)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv

doublingTimes-**ranked(2)**-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-corona-

Table 2 Forecasting performance metrics derived from the weighted and unweighted ensemble models, an auto-regressive integrated moving average model (ARIMA), a generalized additive model (GAM), and simple linear regression model (SLR) based on the daily curve of COVID-19 cases in the USA from 11-May-2020 to 11-June-2020. The weights of the weighted ensemble model are based on relative likelihood. Overall, both ensemble types performed similarly for this forecast, and outperformed the simple statistical models

Model	Forecasting period	MAE	MSE	Coverage 95% PI	WIS
Weighted Ensemble(2)	30	4716.01	30200654.24	66.67	3156.50
Unweighted Ensemble(2)	30	4662.00	29686078.76	66.67	3169.62
Weighted Ensemble(3)	30	5229.91	36934107.63	60.00	3490.51
Unweighted Ensemble(3)	30	5262.52	37441993.33	60.00	3482.98
Weighted Ensemble(4)	30	5023.32	32564946.12	76.67	2926.13
Unweighted Ensemble(4)	30	4836.87	30807937.90	76.67	2942.91
ARIMA	30	7560.80	77139741.86	90.00	4118.39
GAM	30	8345.23	94188590.40	50.00	5466.92
SLR	30	23380.65	583817550.48	0.00	21739.18

```
virus-cases-USA-area-52-05-11-2020.csv
doublingTimes-ranked(3)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-ranked(4)doublingTimes--onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
```

6) Sequence of doubling times of the ensemble models:

```
doublingTimes-Ensemble(2)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-Ensemble(3)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
doublingTimes-Ensemble(4)-onsetfixed-0-typedecline-2-flag1-1-method-0-dist-0-horizon-30-weight_type-1-daily-coronavirus-cases-USA-area-52-05-11-2020.csv
```

We can also compare the performance of unweighted ensemble models by setting the parameter `<weight_type1>=-1` in the options_forecast.m file while the other parameters are kept unchanged. Then we can compare the performance of the unweighted ensemble models (equal weights across top-ranked models) with the weighted ensemble models, where the weights are proportional to the relative likelihood of the models (`<weight_type1>=1`). We can run the function `plotForecast_subepidemicFramework.m` to generate the new set of forecasts with the new models.

The forecasting performance metrics for the weighted and unweighted ensemble models and other statistical time-series models are displayed in Table 2. Overall, the unweighted ensemble models performed similarly as their weighted ensemble counterparts for this forecast and outperformed some popular statistical time-series models such as ARIMA (a brief description of the statistical models is given in Supplementary Text S3).

Conclusion

We have introduced a MATLAB toolbox to fit and forecast time series using the spatial wave sub-epidemic model originally developed to generate short-term forecasts of epidemics [13] and illustrated its functionality using time-series data of the COVID-19 pandemic in the US. In particular, the sub-epidemic model used in this tutorial has shown competitive performance in characterizing and forecasting epidemic trajectories of infectious diseases such as COVID-19, Ebola, and plague [13, 15]. The toolbox can be a helpful resource for policy makers and used as a part of the curriculum for students training in infectious disease modeling, mathematical biology, applied statistics and mathematics, and special courses in epidemic modeling and time-series forecasting.

This new open-source toolbox and associated tutorial will be helpful to a broad community of applied scientists interested in characterizing and forecasting time-series data that results from the aggregation of multiple asynchronous underlying growth processes. Moreover, prior publications have extensively validated the tools presented here [13, 15]. The models and methods included in the toolbox have improved short-term forecasting performance over simpler growth models such as the Richards and generalized-logistic growth models. Moreover, we have ensured publicly available, long-term, and stable hosting of the toolbox in a public GitHub repository. Extensions to the toolbox could include additional components, such as new model features, alternative estimation methods, and additional forecasting performance metrics.

Availability and requirements

Project name: Forecasting growth trajectories using the ensemble spatial wave sub-epidemic modeling framework.

Project home page: https://github.com/gchowell/spatial_wave_subepidemic_framework Operating system(s): Platform independent.

Programming language: MATLAB.

Other requirements: NA.

License: This program is free software: it can be redistributed or modified under the GNU Public License as published by the Free Software Foundation, version 3 of the License.

Any restrictions to use by non-academics: None.

Abbreviations

AIC	Akaike Information Criterion
ARIMA	Auto Regressive Integrated Moving Average
CSSE	Center for System Science and Engineering
CI	Confidence Interval
GLM	Generalized Logistic growth Model
IS	Interval Score
MAE	Mean Absolute Error
MLE	Maximum Likelihood
MSE	Mean Squared Error
ODE	Ordinary Differential Equations
PI	Prediction Interval
USA	United States of America
WIS	Weighted Interval Score

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02241-2.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

None.

Authors' contributions

G.C. conceived and developed the first version of the toolbox and wrote the first draft of the tutorial; G.C., A.T., A.B, S.D., J.M., R.L contributed to analysis and writing subsequent drafts of the tutorial. A.T. produced the tutorial video.

Funding

G.C. is partially supported from NSF grants 2125246 and 2026797 and R01 GM 130900. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Datasets for daily COVID-19 cases reported in the USA are retrieved from the publicly available data tracking system of the Johns Hopkins Center for Systems Science and Engineering (CSSE). Code files can be accessed at https://github.com/gchowell/spatial_wave_subepidemic_framework.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1] Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA. [2] Department of Applied Mathematics, Kyung Hee University, Yongin 17104, Korea. [3] Department of Pediatrics, School of Medicine, Stanford University, Palo Alto, CA, USA. [4] Department of Mathematics, Tulane University, New Orleans, LA, USA.

Received: 22 May 2023 Accepted: 6 May 2024
Published online: 07 June 2024

References

1. Petropoulos F, Apiletti D, Assimakopoulos V, Babai MZ, Barrow DK, Ben Taieb S, Bergmeir C, Bessa RJ, Bijak J, Boylan JE, et al. Forecasting: theory and practice. Int J Forecast. 2022;38(3):705–871.
2. Dimri T, Ahmad S, Sharif M. Time series analysis of climate variables using seasonal ARIMA approach. J Earth Syst Sci. 2020;129:149.
3. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. 2nd ed. OTexts. 2018. p. 384.
4. Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. Int J Sci Eng Appl. 2014;4(2):13.
5. Shamsnia SA, Shahidi N, Liaghat A, Sarraf A, Vahdat SF. Modeling of weather parameters using stochastic methods (ARIMA model)(case study: Abadeh Region, Iran). In: International Conference on Environment and Industrial Innovation. IPCBEE. 2011;12.
6. Tektaş M. Weather forecasting using ANFIS and ARIMA models. Environ Res Eng Manag. 2010;51(1):5–10.
7. Yan P, Chowell G. Quantitative methods for investigating infectious disease outbreaks vol. 70. Cham: Springer; 2019.
8. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model. 2017;2(3):379–98.
9. Chowell G, Castillo-Chavez C, Fenimore PW, Kribs-Zaleta CM, Arriola L, Hyman JM. Model parameters and outbreak control for SARS. Emerg Infect Dis. 2004;10(7):1258.
10. Keeling MJ, Hill EM, Gorsich EE, Penman B, Guyver-Fletcher G, Holmes A, Leng T, McKimm H, Tamborrino M, Dyson L. Predictions of COVID-19 dynamics in the UK: short-term forecasting and analysis of potential exit strategies. PLoS Comput Biol. 2021;17(1):e1008619.
11. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, Zhang Q, Chowell G, Simonsen L, Vespignani A. The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. Epidemics. 2018;22:13–21.
12. Tuncer N, Timsina A, Nuno M, Chowell G, Martcheva M. Parameter identifiability and optimal control of an SARS-CoV-2 model early in the pandemic. J Biol Dyn. 2022;16(1):412–38.

13. Chowell G, Tariq A, Hyman JM. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med. 2019;17(1):164–164.
14. Raimund B, Gerardo C, Leidy Yissedt L-D. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. Math Biosci Eng. 2019;16(5):4250–73.
15. Chowell G, Rothenberg R, Roosa K, Tariq A, Hyman JM, Luo R. Sub-epidemic model forecasts during the first wave of the COVID-19 pandemic in the USA and European hotspots. Mathematics of Public Health. Cham: Springer International Publishing; 2022.
16. Chowell G, Dahal S, Tariq A, Roosa K, Hyman JM, Luo R. An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: application to the COVID-19 pandemic in the USA. PLoS Comput Biol. 2022;18(10): e1010602.
17. Chowell G, Luo R. Ensemble bootstrap methodology for forecasting dynamic growth processes using differential equations: application to epidemic outbreaks. BMC Med Res Methodol. 2021;21(1):34.
18. Banks HT, Hu S, Thompson WC. Modeling and inverse problems in the presence of uncertainty. 1st ed. Chapman and Hall/CRC; 2014. https://doi.org/10.1201/b16760.
19. Roosa K, Luo R, Chowell G. Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study. Math Biosci Eng. 2019;16(5):4299–313.
20. Myung IJ. Tutorial on maximum likelihood estimation. J Math Pyschol. 2003;47:90–100.
21. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Data mining, inference, and prediction. New York: Springer-Verlag New York; 2009.
22. Shanafelt DW, Jones G, Lima M, Perrings C, Chowell G. Forecasting the 2001 foot-and-mouth disease epidemic in the UK. EcoHealth. 2018;15:338–47.
23. Chowell G, Hincapie-Palacio D, Ospina J, Pell B, Tariq A, Dahal S, Moghadas S, Smirnova A, Simonsen L, Viboud C. Using Phenomenological models to characterize transmissibility and Forecast patterns and final Burden of Zika Epidemics. PLoS Curr. 2016;8.
24. Pell B, Kuang Y, Viboud C, Chowell G. Using phenomenological models for forecasting the 2015 Ebola challenge. Epidemics. 2018;22:62–70.
25. Sugiura N. Further analysts of the data by akaike's information criterion and the finite corrections. Commun Stat Theory Methods. 1978;7:13–26.
26. Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. Biometrika. 1989;76:297–307.
27. Gneiting T, Raftery AE. Strictly proper Scoring rules, Prediction, and estimation. J Am Stat Assoc. 2007;102(477):359–78.
28. Kuhn M, Johnson K. Applied predictive modeling, vol. 26. New York: Springer; 2013.
29. Competitor's Guide: Prizes and Rules.[https://www.m4.unic.ac.cy/wp-content/uploads/2018/03/M4-Competitors-Guide.pdf].
30. Tariq A, Chakhaia T, Dahal S, Ewing A, Hua X, Ofori SK, Prince O, Salindri AD, Adeniyi AE, Banda JM, et al. An investigation of spatial-temporal patterns and predictions of the coronavirus 2019 pandemic in Colombia, 2020–2021. PLoS Negl Trop Dis. 2022;16(3): e0010228.
31. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. PLoS Comput Biol. 2021;17(2): e1008618.
32. Hwang E. Prediction intervals of the COVID-19 cases by HAR models with growth rates and vaccination rates in top eight affected countries: bootstrap improvement. Chaos Solitons Fractals. 2022;155:111789–111789.
33. Roosa K, Tariq A, Yan P, Hyman JM, Chowell G. Multi-model forecasts of the ongoing ebola epidemic in the Democratic Republic of Congo, March 2013-October 2019. J R Soc Interface. 2020;17(169):20200447.
34. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, Gerding A, Gneiting T, House KH, Huang Y, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proc Natl Acad Sci U S A. 2022;119(15): e2113561119.
35. Muniz-Rodriguez K, Chowell G, Cheung CH, Jia D, Lai PY, Lee Y, Liu M, Ofori SK, Roosa KM, Simonsen L, et al. Doubling time of the COVID-19 epidemic by Province, China. Emerg Infect Dis. 2020;26(8):1912–4.
36. Smirnova A, DeCamp L, Chowell G. Mathematical and statistical analysis of doubling times to investigate the early spread of epidemics: application to the COVID-19 pandemic. Mathematics. 2021;9(6): 625.
37. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B: Biol Sci. 2007;274(1609):599–604.
38. Chowell G, Luo R, Sun K, Roosa K, Tariq A, Viboud C. Real-time forecasting of epidemic trajectories using computational dynamic ensembles. Epidemics. 2020;30: 100379.
39. Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. PLoS Comput Biol. 2018;14(2): e1005910.
40. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer; 2002. p. 488.
41. Hopkins J. CSSE Covid-19 timeseries. GitHub; 2022.
42. Tariq A. GitHub Repository. 2022.
43. Roosa K, Chowell G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. Theor Biol Med Model. 2019;16(1):1.
44. Chowell G, Tariq A, Hyman JM. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med. 2019;17(1):164.

Publisher's Note