Title: An ensemble $n$-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the COVID-19 pandemic in the USA

Short title: An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemics

Gerardo Chowell$^{1,2}$, Sushma Dahal$^1$, Amna Tariq$^1$, Kimberlyn Roosa$^3$, James M. Hyman$^4$, Ruiyan Luo$^1$

$^1$ Department of Population Health Sciences, School of Public Health, Georgia State University, Atlanta, GA, USA.
$^2$ Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA.
$^3$ National Institute for Mathematical and Biological Synthesis (NIMBioS), University of Tennessee, Knoxville, TN, USA
$^4$ Department of Mathematics, Center for Computational Science, Tulane University, New Orleans, LA, USA.

* Corresponding author (gchowell@gsu.edu)

Abstract

We analyze an ensemble of $n$-sub-epidemic modeling for forecasting the trajectory of epidemics and pandemics. These ensemble modeling approaches, and models that integrate sub-epidemics to capture complex temporal dynamics, have demonstrated powerful forecasting capability. This modeling framework can characterize complex epidemic patterns, including plateaus, epidemic resurgences, and epidemic waves characterized by multiple peaks of different sizes. We systematically assess their calibration and short-term forecasting performance in short-term forecasts for the COVID-19 pandemic in the USA from late April 2020 to late February 2022. We compare their performance with two commonly used statistical ARIMA models. The best fit
sub-epidemic model and three ensemble models constructed using the top-ranking sub-epidemic
models consistently outperformed the ARIMA models in terms of the weighted interval score
(WIS) and the coverage of the 95% prediction interval across the 10-, 20-, and 30-day short-term
forecasts. In the 30-day forecasts, the average WIS ranged from 377.6 to 421.3 for the sub-
epidemic models, whereas it ranged from 439.29 to 767.05 for the ARIMA models. Across 98
short-term forecasts, the ensemble model incorporating the top four ranking sub-epidemic
models (Ensemble(4)) outperformed the (log) ARIMA model 66.3% of the time, and the
ARIMA model 69.4% of the time in 30-day ahead forecasts in terms of the WIS. Ensemble(4)
consistently yielded the best performance in terms of the metrics that account for the uncertainty
of the predictions. This framework could be readily applied to investigate the spread of
epidemics and pandemics beyond COVID-19, as well as other dynamic growth processes found
in nature and society that would benefit from short-term predictions.

Summary
The COVID-19 pandemic has highlighted the urgent need to develop reliable tools to forecast
the trajectory of epidemics and pandemics in near real-time. We describe and apply an ensemble
n-sub-epidemic modeling framework for forecasting the trajectory of epidemics and pandemics.
We systematically assess its calibration and short-term forecasting performance in weekly 10-30
days ahead forecasts for the COVID-19 pandemic in the USA from late April 2020 to late
February 2022 and compare its performance with two different statistical ARIMA models. This
framework demonstrated reliable forecasting performance and substantially outcompeted the
ARIMA models. The forecasting performance was consistently best for the ensemble sub-
epidemic models incorporating a higher number of top-ranking sub-epidemic models. The
ensemble model incorporating the top four ranking sub-epidemic models consistently yielded the
best performance, particularly in terms of the coverage rate of the 95% prediction interval and the weighted interval score. This framework can be applied to forecast other growth processes found in nature and society including the spread of information through social media.
Introduction

The coronavirus disease 2019 (COVID-19) pandemic has amplified the critical need for reliable tools to forecast the trajectory of epidemics and pandemics in near real-time. During the early stages of the COVID-19 pandemic, multiple modeling teams embarked on the challenging task of producing short-term forecasts of the course of the COVID-19 pandemic in terms of the trajectory for the number of new cases, hospitalizations, or deaths (e.g., [1-10]). Soon after the epidemic started, our research team published short-term forecasts of the pandemic during the early outbreaks of the novel coronavirus in China [4] and subsequently focused on producing weekly forecasts for the USA [11]. In a related effort, the US COVID-19 Forecasting Hub brought together multiple research teams to synthesize weekly short-term forecasts of the COVID-19 pandemic in the USA [12]. It is time to systematically and rigorously evaluate the forecasting performance of these different pandemic forecasting efforts and document the lessons learned to continue advancing our understanding of epidemic forecasting.

Ensemble modeling approaches and models that integrate sub-epidemics to capture complex temporal dynamics have demonstrated powerful forecasting capability (e.g., [13] [14-17]). In prior work, we developed a sub-epidemic modeling framework to characterize and improve forecasting accuracy during complex epidemic waves [13]. This mathematical framework characterizes epidemic curves by aggregating multiple asynchronous sub-epidemics and outperforms simpler growth models at providing short-term forecasts of various infectious disease outbreaks [13, 18]. It is possible to model sub-epidemics associated with transmission chains that are asynchronously triggered and progress somewhat independently from the other
sub-epidemics. This framework supports a family of sub-epidemic models that yield similar fits to the calibration data, but their corresponding forecasts could produce diverging trajectories.

Ensemble modeling aims to boost forecasting performance by systematically integrating the predictive accuracy tied to individual models [16, 19-21]. Past work indicates that multimodel ensemble approaches are powerful forecasting tools that frequently outperform individual models in epidemic forecasts [14, 15, 22-27]. We extend prior sub-epidemic modeling work and propose an ensemble sub-epidemic modeling framework for forecasting the trajectory of epidemics and pandemics. In this model, the sub-epidemics can start at different time points and may follow different growth rates, scaling of growth, and final sizes. Hence, this ensemble modeling framework can characterize more diverse epidemic patterns which were impossible to capture by earlier sub-epidemic models, including plateaus, epidemic resurgences, and epidemic waves characterized by multiple peaks of different sizes.

Here, we systematically assess the calibration and short-term forecasting performance in weekly 10-30 day forecasts in the context of the COVID-19 pandemic in the USA from late April 2020 to late February 2022, including the Omicron-dominated wave. We then compare the performance of the ensemble modeling framework with a set of Autoregressive Integrated Moving Average (ARIMA) models, following the EPIFORGE 2020 guidelines to report epidemic forecasts [28]. Our extended ensemble modeling framework substantially outperforms individual top-ranking sub-epidemic models and the ARIMA models based on standard performance metrics that account for the uncertainty of the predictions.

Data
We used daily COVID-19 deaths reported in the USA from the publicly available data tracking system of the Johns Hopkins Center for Systems Science and Engineering (CSSE) from 27 February 2020 to 30 March 2022 [29]. The data is updated on the CSSE webpage once every day at 23:59 (UTC) and is read from the daily case report. The data is also publicly available in the GitHub repository [30].

**n-sub-epidemic model**

We model epidemic trajectories comprised of one or more overlapping and asynchronous sub-epidemics. That is, the sub-epidemics are used as building blocks to characterize more complex epidemic trajectories. The mathematical equation for the sub-epidemic building block is the 3-parameter generalized-logistic growth model (GLM), which has performed well in short-term forecasts of single outbreak trajectories for different infectious diseases, including COVID-19 [31-33]. This model is given by the following differential equation:

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

where \(\frac{dC(t)}{dt}\) describes the curve of daily deaths over time \(t\). The cumulative curve at time \(t\) is given by \(C(t)\), while \(r\) is a positive parameter denoting 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 number of new deaths 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.

An \( n \)-sub-epidemic trajectory comprises \( n \) overlapping sub-epidemics and is given by the following system of coupled differential equations:

\[
\frac{dC_i(t)}{dt} = C_i'(t) = A_i(t) r_i C_i^{p_i}(t) \left( 1 - \frac{C_i(t)}{K_{0i}} \right),
\]

Where \( C_i(t) \) tracks the cumulative number of deaths for sub-epidemic \( i \), and the parameters that characterize the shape of the \( i \)th sub-epidemic are given by \( (r_i, p_i, K_{0i}) \), for \( i = 1, ..., n \). Thus, the 1-sub-epidemic model is equivalent to the generalized growth model described above. When \( n > 1 \), we model the onset timing of the \((i + 1)\)th sub-epidemic, where \((i + 1) \leq n \), by employing an indicator variable given by \( A_i(t) \) so that the \((i + 1)\)th sub-epidemic is triggered when the cumulative curve of the \( i \)th sub-epidemic exceeds \( C_{thr} \).

The \((i + 1)\)th sub-epidemic is only triggered when \( C_{thr} \leq K_{0i} \). Hence, we have:

\[
A_i(t) = \begin{cases} 
1, & C_{i-1}(t) > C_{thr} \\
0, & \text{Otherwise} 
\end{cases} \quad i = 2, ..., n,
\]

where \( A_1(t) = 1 \) for the first sub-epidemic. Hence, the total number of parameters that are needed to model an \( n \)-sub-epidemic trajectory is given by \( 3n + 1 \). The initial number of deaths is given
by $C_1(0) = I_0$, where $I_0$ is the initial number of deaths in the observed data. The cumulative
curve of the $n$-sub-epidemic trajectory is given by:

$$C_{tot}(t) = \sum_{i=1}^{n} C_i(t).$$

The $n$-sub-epidemic wave model can characterize diverse epidemic patterns, including epidemic
plateaus where the epidemic stabilizes at a high level for an extended period, epidemic
resurgences where the number of cases increases again after a low incidence period, and
epidemic waves characterized by multiple peaks.

**Parameter estimation**

To reduce the noise in the original data due to artificial reasons such as the weekend effects, we
use the 7-day moving average of daily death series to fit the $n$-sub-epidemic model. Let

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

denote the smoothed daily COVID-19 death 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 $y_{t_j}$, $j=1,2, ..., n_d$, is the average of the death counts at the neighboring
seven days ($t_{j-3}, t_{j-2}, t_{j-1}, t_j, t_{j+1}, t_{j+2}, t_{j+3}$). We will use this smoothed data to estimate a total
of $3n + 1$ model parameters, namely $\Theta = (C_{thr}, r_1, p_1, K_{0_1}, ..., r_n, p_n, K_{0_n})$. Let $f(t, \Theta)$ denote
the expected curve of new COVID-19 deaths of the epidemic’s trajectory. We can estimate
model parameters by fitting the model solution to the observed data via nonlinear least squares
[34] or via maximum likelihood estimation assuming a specific error structure [35]. For
nonlinear least squares, this is achieved by searching for the set of parameters \( \hat{\theta} \) that minimizes the sum of squared differences between the observed data \( y_{t_1}, y_{t_2}, \ldots, y_{t_{n_d}} \) and the model mean corresponds to \( f(t, \theta) \). That is, \( \theta = (C_{th}, r_1, p_1, K_{0_1}, \ldots, r_n, p_n, K_{0_n}) \) is estimated by \( \hat{\theta} = \arg \min \sum_{j=1}^{n_d} (f(t_j, \theta) - y_{t_j})^2 \).

This parameter estimation method weights each of the data points equally and does not 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 (e.g., number of deaths) denoted by \( f(t, \theta) \) at time \( t \) [36]. This method yields asymptotically unbiased point estimates regardless of any misspecification of the variance-covariance error structure. Hence, the estimated model mean \( f(t_i, \hat{\theta}) \) yields the best fit to observed data \( y_{t_i} \) in terms of squared L2 norm. In Matlab, we can use the \textit{fmincon} function to set the optimization problem.

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 [37]. We generate \( B \) bootstrap samples from the best-fit model \( f(t, \hat{\theta}) \), with an assumed error structure, to quantify the uncertainty of the parameter estimates and construct confidence intervals. Typically, the error structure in the data is modelled using a probability model such as the Poisson or negative binomial distribution. Because the time-series data we are fitting to involve large counts, the Poisson or negative binomial distribution can be well approximated by a normal distribution for large numbers. So, using the best-fit model \( f(t, \hat{\theta}) \), we generate \( B \)-times replicated simulated datasets of size \( n_d \), where the observation at time \( t_j \) is sampled from a
normal distribution with mean $f(t_j, \bar{\Theta})$ and variance $\frac{\sum_{j=1}^{n_d} (f(t_j, \bar{\Theta}) - y_{t_j})^2}{n_d - (3n + 1)}$. Next, we refit the model to each of the $B$ simulated datasets to re-estimate parameters for each. The new parameter estimates for each realization are denoted by $\bar{\Theta}_b$ where $b = 1, 2, ..., B$. Using the sets of re-estimated parameters ($\bar{\Theta}_b$), 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 $f(t, \bar{\Theta}_1), f(t, \bar{\Theta}_2), ..., f(t, \bar{\Theta}_B)$.

Model-based forecasts with quantified uncertainty

Forecasting the model $f(t, \bar{\Theta})$, $h$ days ahead provides an estimate for $f(t + h, \bar{\Theta})$. The uncertainty of the forecasted value can be obtained using the previously described parametric bootstrap method. Let

$$f(t + h, \bar{\Theta}_1), f(t + h, \bar{\Theta}_2), ..., f(t + h, \bar{\Theta}_B)$$

denote the forecasted value of the current state of the system propagated by a horizon of $h$ time units, where $\bar{\Theta}_b$ denotes the estimation of parameter set $\Theta$ from the $b_{th}$ bootstrap sample. We can use these values to calculate the bootstrap variance as the measure of the uncertainty of the forecasts and use the 2.5% and 97.5% percentiles to construct the 95% prediction intervals (PI).

Model selection

We considered a set of $n$-sub-epidemic models where $1 \leq n \leq 2$ and ranked them from best to worst according to the $AIC_c$ which is given by [38, 39]:

CC-BY 4.0 International license
It is made available under a CC-BY 4.0 International license
\[ AIC_c = n_d \log(SSE) + 2m + \frac{2m(m + 1)}{n_d - m - 1} \]

where \( SSE = \sum_{j=1}^{n_d} (f(t_j, \bar{\theta}) - y_{t_j})^2 \), \( m = 3n + 1 \) is the number of model parameters, and \( n_d \) is the number of data points. The \( AIC_c \) for the parameter estimation from the nonlinear least-squares fit, which implicitly assumes normal distribution for error.

We selected the four top ranking sub-epidemic models for further analyses. We used them to construct three ensemble sub-epidemic models, which we refer to as: Ensemble(2), Ensemble(3), and Ensemble(4). The next section describes the process of constructing these ensemble models from the top-ranking sub-epidemic models.

**Constructing Ensemble Models from top-ranking models**

Ensemble models that combine the strength of multiple models may exhibit significantly enhanced predictive performance (e.g., [14-17]). Here we generate ensemble models from the weighted combination of the highest-ranking sub-epidemic models as deemed by the \( AIC_{c_i} \) for the \( i \)-th model where \( AIC_{c_1} \leq \cdots \leq AIC_{c_I} \) and \( i = 1, \ldots, I \). An ensemble derived from the top-ranking \( I \) models is denoted by Ensemble(\( I \)) and illustrated in Figure 1. Thus, Ensemble(2) and Ensemble(3) refer to the ensemble models generated from the weighted combination of the top-ranking 2 and 3 models, respectively. We compute the weight \( w_i \) for the \( i \)-th model, \( i = 1, \ldots, I \), where \( \sum w_i = 1 \) as follows:
\[ w_i = \frac{1}{\frac{1}{\text{AIC}_{c_1}} + \frac{1}{\text{AIC}_{c_2}} + \cdots + \frac{1}{\text{AIC}_{c_I}}} \] for all \( i = 1, 2, \ldots, I \),

and hence \( w_I \leq \cdots \leq w_1 \).

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

\[ f_{\text{ens}(I)}(t) = \sum_{i=1}^{I} w_i f_i(t, \tilde{\Theta}^{(i)}) \]

where given the training data, \( \tilde{\Theta}^{(i)} \) denotes the set of estimated parameters, and \( f_i(t, \tilde{\Theta}^{(i)}) \) denotes the estimated mean curve of daily COVID-19 deaths, 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 [16]. Our MATLAB (The Mathworks, Inc) code for model fitting and forecasting is publicly available in the GitHub repository [30].

**Figure 1.** 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_i} for the \( i \)-th model where \( \text{AIC}_{c_1} \leq \cdots \leq \text{AIC}_{c_I} \) and \( i = 1, \ldots, I \). An ensemble derived from the top-ranking \( I \) models is denoted by Ensemble(I).

As a sensitivity analysis, we also investigated how the ensemble sub-epidemic models performed when the ensemble weights were proportional to the relative likelihood (\( l \)) rather than the reciprocal of the AIC. Let AIC_{min} denote the minimum AIC from the set of models. The relative
likelihood of model $i$ is given by $l_i = e^{((AIC_{\text{min}} - AIC_i)/2)}$ [40]. We compute the weight $w_i$ for the $i$-th model where $\sum w_i = 1$ as follows:

$$w_i = \frac{l_i}{\sum l_i} \quad \text{for all } i = 1, 2, \ldots, I,$$

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

**Auto-regressive integrated moving average models (ARIMA)**

We also generated short-term predictions of the pandemic trajectory using ARIMA models to compare their performance with that of the sub-epidemic models. ARIMA models have been frequently employed to forecast trends in finance [41-43] and weather [44-46]. The ARIMA ($p, d, q$) process is given by

$$\phi(B)(1 - B)^d y_t = c + \theta(B)\epsilon_t$$

or equivalently as $\phi(B)(1 - B)^d (y_t - \mu t^d/d!) = \theta(B)\epsilon_t$, where $p$ is the order of the AR model, $d$ is the degree of differencing, $q$ is the order of the MA model, $\{\epsilon_t\}$ is a white noise process with mean 0 and variance $\sigma^2$, and $B$ denotes the backshift operator. The $p$-order polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and the $q$-order polynomial $d \theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$ are assumed to have no roots inside the unit circle to ensure causality and invertibility. The constant $c = \mu (1 - \phi_1 - \cdots - \phi_p)$, and $\mu$ is the mean of $(1 - B)^d y_t$. When $d=0$, $\mu$ is the mean of $y_t$. 

The `auto.arima` function in the R package “forecast” is used to select orders and build the model [47]. First, the degree of differencing $0 \leq d \leq 2$ is selected based on successive KPSS unit-root tests [48], which test the data for a unit root; if the test result is significant, the differenced data is tested for a unit root; and this procedure is repeated until the first insignificant result is obtained. Then given $d$, the orders $p$ and $q$ are selected based on the AIC$_c$ for the $d$-times differenced data. For $d=0$ or $d=1$, a constant will be included if it improves the AIC$_c$ value; for $d>1$, the constant $\mu$ is fixed at 0 to avoid the model having a quadratic or higher order trend, which is dangerous when forecasting. The final model is fitted using the maximum likelihood estimation.

To guarantee the forecasted values and prediction intervals are above zero, we take the following two strategies. In the first one, we conduct the ARIMA order selection and model fitting using the log-transformed data. Then we take the exponential of the forecasted values and the PI bounds to predict the incident death counts and get the PIs. We refer to this approach as the (log) ARIMA throughout the manuscript. In the second case, the negative values are set as zero. Then, it is possible that the actual coverage probability of such PIs can be smaller than the nominal value (95%). We refer to this approach as ARIMA throughout the manuscript.

**Forecasting strategy and performance metrics**

We conducted short-term forecasts using the top-ranking $n$-sub-epidemic model $(1 \leq n \leq 2)$ and three ensemble models constructed with the top-ranking sub-epidemic models namely Ensemble(2), Ensemble(3), and Ensemble(4). For comparison, we also generated short-term
forecasts using the previously described ARIMA models. Overall, we conducted 588 forecasts across models.

Using a 90-day calibration period for each model, we conducted a total of 98 weekly sequential 10-day, 20-day and 30-day forecasts from 20 April 2020 to 28 February 2022, spanning five pandemic waves. This range of forecasting horizons is comparable to that investigated in prior COVID-19 forecasting studies [49]. This period covers the latter part of the early spring wave, a summer wave in 2020, a fall-winter 2020/2021 wave, the summer-fall wave in 2021, and the winter 2022 wave.

To assess the forecasting performance, we used four performance metrics: the mean absolute error (MAE), the mean squared error (MSE), the coverage of the 95% prediction intervals, and the mean interval score (MIS) [50]. The mean absolute error (MAE) is given by:

$$\text{MAE} = \frac{1}{N} \sum_{h=1}^{N} |f(t_h, \hat{\theta}) - \bar{y}_{t_h}|.$$  

Here $\bar{y}_{t_h}$ is the time series of the original death counts (unsmoothed) of the $h$-time units ahead forecasts, where $t_h$ are the time points of the time series data [51]. Similarly, the mean squared error (MSE) is given by:

$$\text{MSE} = \frac{1}{N} \sum_{h=1}^{N} (f(t_h, \hat{\theta}) - \bar{y}_{t_h})^2.$$
We also employed two metrics that account for prediction uncertainty: the **coverage rate of the 95% PI** e.g., the proportion of the observations that fall within the 95% PI as well as the **weighted interval score** (WIS) [50, 52] which is a proper score. The WIS and the coverage rate of the 95% PIs take into account the uncertainty of the predictions, whereas the MAE and MSE only assess the closeness of the mean trajectory of the epidemic to the observations [53].

Recent epidemic forecasting studies have embraced the Interval Score (IS) for quantifying model forecasting performance [18, 24, 49, 54]. The WIS provides quantiles of predictive forecast distribution by combining a set of ISs for probabilistic forecasts. An IS is a simple proper score that requires only a central \((1-\alpha)\times100\%\) PI [50] and is described as

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

In this equation 1 refers to the indicator function, meaning that \(1(y < l) = 1\) 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 \(\frac{2}{\alpha} \times (l - y) \times 1(y < l)\) 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 $\alpha$.

3. An analogous penalty term $\frac{2}{\alpha} \times (y - u) \times 1(y > u)$ 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 $(1 - \alpha_1) < (1 - \alpha_2) < \ldots < (1 - \alpha_K)$ along with the predictive median, $m$, 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:

$$WIS_{\alpha;K}(F, y) = \frac{1}{K + \frac{1}{2}} \cdot (w_0 \cdot |y - m| + \sum_{k=1}^{K} w_k \cdot IS_{\alpha_k}(F, y))$$

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 entire distribution is to the observation in units on the scale of the observed data [10, 55].

Results

Quality of the sub-epidemic model fits

The best fit sub-epidemic model and three ensemble models constructed using the top-ranking sub-epidemic models (Ensemble(2), Ensemble(3), Ensemble(4)) yielded similar quality fits to 98
sequential weekly calibration periods from 20-April-2020 to 28-February-2022 (Figure 2, Table 1). For instance, the average WIS was ~247 with little variation across models (Table 1). The coverage rate of the 95% PIs averaged 97% and ranged from 91% to 100% during the study period. Moreover, all performance metrics displayed similar temporal trends (Figure 2).

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean absolute error (MSE)</th>
<th>Mean squared error (MAE)</th>
<th>Percentage coverage of the 95% prediction interval</th>
<th>Weighted Interval Score (WIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best fit sub-epidemic model</td>
<td>309260.00</td>
<td>394.74</td>
<td>97.06</td>
<td>247.28</td>
</tr>
<tr>
<td>Ensemble(2) model</td>
<td>308300.00</td>
<td>394.91</td>
<td>97.30</td>
<td>246.93</td>
</tr>
<tr>
<td>Ensemble(3) model</td>
<td>308620.00</td>
<td>395.24</td>
<td>97.46</td>
<td>247.09</td>
</tr>
<tr>
<td>Ensemble(4) model</td>
<td>309160.00</td>
<td>396.17</td>
<td>97.46</td>
<td>247.33</td>
</tr>
</tbody>
</table>

*The Ensemble(i) model incorporates the top i ranked sub-epidemic models in the ensemble as described in the text.

Table 1. Mean performance metrics quantifying the quality of model fits across 98 sequential weekly calibration periods of the daily time series of COVID-19 deaths in the USA from 20-April-2020 through 22-February 2022.

Figure 2. Performance metrics quantifying the quality of sub-epidemic model fits to 98 sequential weekly calibration periods of the daily time series of COVID-19 deaths in the USA from 20-April-2020 through 22-February 2022. The best fit sub-epidemic model and three ensemble models constructed using the top-ranking sub-epidemic models (Ensemble(2), Ensemble(3), Ensemble(4)) yielded similar quality fits.
Representative fits of the top-ranking sub-epidemic models to the daily curve of COVID-19 deaths in the USA from 27-Feb-2020 to 20-April-2020 are shown in Figure 3. Although these sub-epidemic models fit the data well, each of them results from the aggregation of two sub-epidemics characterized by different growth rates, scaling of growth, and outbreak sizes as shown in Figure 4.

Figure 3. Representative fits of the top-ranking sub-epidemic models to the daily curve of COVID-19 deaths in the USA from 27-Feb-2020 to 20-April-2020. The sub-epidemic models capture well the entire epidemic curve, including the latter plateau dynamics, by considering models with two sub-epidemics. The best model fit (solid red line) and 95% prediction interval (dashed red lines) are shown in the left panels. The cyan curves correspond to the associated uncertainty from individual bootstrapped curves. The sub-epidemic profiles are shown in the center panels, where the red and blue curves represent the two sub-epidemics and the grey curves are the estimated epidemic trajectories. For each model fit, the residuals are also shown (right panels). Black circles correspond to the data points.

Figure 4. Parameter estimates for the first (top panel) and the second sub-epidemics (bottom panels) were derived for the top-ranking sub-epidemic model after fitting the sub-epidemic modeling framework to the daily curve of COVID-19 deaths in the USA from 27-Feb-2020 to 20-April-2020 (see also Figure 2). Parameter estimates for both sub-epidemics are well identified, as indicated by their relatively narrow bootstrap confidence intervals.

Short-term forecasting performance
The best fit sub-epidemic model and three ensemble models constructed using the top-ranking sub-epidemic models (Ensemble(2), Ensemble(3), Ensemble(4)) consistently outperformed the ARIMA models in terms of the weighted interval score (WIS) and the coverage of the 95% prediction interval across the 10, 20 and 30 day short-term forecasts (Table 2). For instance, for 30-day forecasts, the average WIS ranged from 377.6 to 421.3 for the sub-epidemic models, whereas, it ranged from 439.29 to 767.05 for the ARIMA models. Across 98 short-term forecasts, the Ensemble(4) outperformed the (log) ARIMA model 66.3% of the time and the ARIMA model 69.4% of the time in 30-day ahead forecasts in terms of the WIS (Figure 5 & Figure 6). Similarly, the coverage of the 95% PI ranged from 82.2% to 88.2% for the sub-epidemic models, whereas it ranged from 58% to 60.3% for the ARIMA models in 30-day forecasts. In terms of the coverage of the 95% PI, the Ensemble(4) outperformed the (log) ARIMA model 89.8% of the time and the ARIMA model 91.8% of the time (Figure 5 & Figure 6). Forecasting performance generally improved as the number of top-ranking sub-epidemic models included in the ensemble increased (Table 1). The Ensemble(4) model consistently yielded the best performance in terms of the metrics that account for the uncertainty of the predictions.

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean absolute error (MSE)</th>
<th>Mean squared error (MAE)</th>
<th>Percentage coverage of the 95% prediction interval</th>
<th>Weighted Interval Score (WIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 days ahead</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-ranked sub-epidemic model</td>
<td>551740.00</td>
<td>535.16</td>
<td>87.14</td>
<td>352.00</td>
</tr>
<tr>
<td>Ensemble(2) model</td>
<td>504560.00</td>
<td>516.44</td>
<td>88.88</td>
<td>331.83</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Ensemble(3) model</th>
<th>491020.00</th>
<th>513.39</th>
<th>89.29</th>
<th>328.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble(4) model</td>
<td>491740.00</td>
<td>513.14</td>
<td>89.39</td>
<td>326.56</td>
</tr>
<tr>
<td>(log) ARIMA model</td>
<td>424880.00</td>
<td>458.72</td>
<td>42.45</td>
<td>365.19</td>
</tr>
<tr>
<td>ARIMA model</td>
<td>430070.00</td>
<td>467.18</td>
<td>43.06</td>
<td>380.47</td>
</tr>
</tbody>
</table>

**20 days ahead**

<table>
<thead>
<tr>
<th>Top-ranked sub-epidemic model</th>
<th>646880.00</th>
<th>570.34</th>
<th>85.15</th>
<th>382.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble(2) model</td>
<td>576700.00</td>
<td>544.35</td>
<td>88.57</td>
<td>354.04</td>
</tr>
<tr>
<td>Ensemble(3) model</td>
<td>558890.00</td>
<td>540.71</td>
<td><strong>89.59</strong></td>
<td>350.73</td>
</tr>
<tr>
<td>Ensemble(4) model</td>
<td>557130.00</td>
<td>539.30</td>
<td><strong>89.44</strong></td>
<td><strong>346.83</strong></td>
</tr>
<tr>
<td>(log) ARIMA model</td>
<td>591980.00</td>
<td>536.22</td>
<td>51.07</td>
<td>422.41</td>
</tr>
<tr>
<td>ARIMA model</td>
<td><strong>538690.00</strong></td>
<td><strong>528.87</strong></td>
<td>55.05</td>
<td>404.92</td>
</tr>
</tbody>
</table>

**30 days ahead**

<table>
<thead>
<tr>
<th>Top-ranked sub-epidemic model</th>
<th>749560.00</th>
<th>613.75</th>
<th>82.18</th>
<th>421.29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble(2) model</td>
<td>670740.00</td>
<td>586.52</td>
<td>87.35</td>
<td>383.36</td>
</tr>
<tr>
<td>Ensemble(3) model</td>
<td>650790.00</td>
<td>584.20</td>
<td><strong>88.20</strong></td>
<td>382.79</td>
</tr>
<tr>
<td>Ensemble(4) model</td>
<td><strong>644270.00</strong></td>
<td><strong>579.77</strong></td>
<td><strong>88.16</strong></td>
<td><strong>377.64</strong></td>
</tr>
<tr>
<td>(log) ARIMA model</td>
<td>818530.00</td>
<td>621.58</td>
<td>57.99</td>
<td>767.05</td>
</tr>
<tr>
<td>ARIMA model</td>
<td>656480.00</td>
<td>591.93</td>
<td>60.34</td>
<td>439.29</td>
</tr>
</tbody>
</table>

*The Ensemble(i) model incorporates the top i ranked sub-epidemic models in the ensemble as described in the text.

**Table 2.** Mean forecasting performance metrics for the sub-epidemic models (ensemble weights are proportional to the reciprocal of the AICc) and the ARIMA models across 98 sequential weekly calibration periods of the daily time series of COVID-19 deaths in the USA from 20-April-2020 through 22-February 2022. Values highlighted in bold correspond to the best performance metrics.
Figure 5. Forecasting performance metrics for the (log) ARIMA model and the Ensemble(4) model across 98 30-day forecasts. The symbol (^) indicates weekly forecasts where the Ensemble(4) model outperformed the (log) ARIMA model. For example, the Ensemble(4) outperformed the (log) ARIMA model 66.3% of the time in terms of the WIS and 89.8% of the time in terms of the coverage rate of the 95% PI (Figure 4 & Figure 6).

Figure 6. Forecasting performance metrics for the ARIMA model and the Ensemble(4) model across 98 30-day forecasts. The symbol (^) indicates weekly forecasts where the Ensemble(4) model outperforms the ARIMA model. For instance, the Ensemble(4) outperformed the ARIMA model 69.4% of the time in terms of the WIS and 91.8% of the time in terms of the coverage rate of the 95% PI (Figure 4 & Figure 6).

In terms of the metrics based on point estimate information, the ARIMA models showed lower overall MSE or MAE compared to the sub-epidemic models in 10 and 20-day forecasts, but the Ensemble(4) achieved the best forecasting performance in 30-day forecasts (Table 2). Overall, the forecasting performance deteriorated at longer forecasting horizons across all models considered in our study.

Representative 30-day forecasts of the top-ranking sub-epidemic models to the daily curve of COVID-19 deaths in the USA from 20-April-2020 to 20-May-2022 are shown in Figure 7. The corresponding sub-epidemic profiles of the forecasts are shown in Figure 8. These models support forecasts with diverging trajectories even though they yield similar fits to the calibration.
period. For instance, the top-ranked sub-epidemic model predicts a decline in the mortality curve, whereas the second-ranked model predicts a stable pattern during the next 30 days (Figure 7). The corresponding forecasts generated from three ensemble models (Ensemble(2), Ensemble(3), Ensemble(4)) built from the top-ranking sub-epidemic models are shown in Figure 9. The individual 30-day ahead predictions across 98 forecasting periods generated by the Ensemble(4) and the ARIMA models are available in the GitHub repository [30].

Figure 7. Representative 30-day forecasts of the top-ranking sub-epidemic models to the daily curve of COVID-19 deaths in the USA from 20-April-2020 to 20-May-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. Circles correspond to the data points. These four top-ranking models support forecasts with diverging trajectories even though they yield similar fits to the calibration period. For instance, the 1st ranked sub-epidemic model predicts a decline in the mortality curve whereas the 2nd ranked model predicts a stable pattern during the next 30 days.

Figure 8. Representative sub-epidemic profiles of the forecasts derived from the top-ranking sub-epidemic models to the daily curve of COVID-19 deaths in the USA from 20-April-2020 to 20-May-2022. The model fit (solid line) and 95% prediction interval (shaded area) are also shown. Black circles correspond to the calibration data. Blue and red curves represent different sub-epidemics of the epidemic wave profile. Gray curves correspond to the overall epidemic trajectory obtained by aggregating the sub-epidemic curves. The vertical line indicates the start time of the forecast.
Figure 9. Representative sub-epidemic ensemble model forecasts (Ensemble(2), Ensemble(3), Ensemble(4)) of COVID-19 deaths in the USA from 20-April-2020 to 20-May-2022. Circles correspond to the data points. The model fits (solid line) and 95% prediction intervals (shaded area) are shown. Circles correspond to the data points. The vertical line indicates the start time of the forecast.

In sensitivity analyses, defining ensemble weights as proportional to the relative likelihood did not achieve better performance relative to the ensemble models generated using weights proportional to the reciprocal of the AICc. Moreover, the rank of the ensemble models was not affected by the type of weights (Table 3).

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean absolute error (MSE)</th>
<th>Mean squared error (MAE)</th>
<th>Percentage coverage of the 95% prediction interval</th>
<th>Weighted Interval Score (WIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10 days ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-ranked sub-epidemic model</td>
<td>551740.00</td>
<td>535.16</td>
<td>87.14</td>
<td>352.00</td>
</tr>
<tr>
<td>Ensemble(2) model</td>
<td>548540.00</td>
<td>534.14</td>
<td><strong>87.25</strong></td>
<td>348.66</td>
</tr>
<tr>
<td>Ensemble(3) model</td>
<td>547220.00</td>
<td>533.51</td>
<td><strong>87.25</strong></td>
<td>347.99</td>
</tr>
<tr>
<td>Ensemble(4) model</td>
<td>546350.00</td>
<td>533.23</td>
<td><strong>87.35</strong></td>
<td>347.60</td>
</tr>
<tr>
<td>(log) ARIMA model</td>
<td><strong>424880.00</strong></td>
<td><strong>458.72</strong></td>
<td>42.45</td>
<td>365.19</td>
</tr>
<tr>
<td>ARIMA model</td>
<td>430070.00</td>
<td>467.18</td>
<td>43.06</td>
<td>380.47</td>
</tr>
<tr>
<td><strong>20 days ahead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-ranked sub-epidemic model</td>
<td>646880.00</td>
<td>570.34</td>
<td>85.15</td>
<td>382.90</td>
</tr>
<tr>
<td>Ensemble(2) model</td>
<td>640240.00</td>
<td>567.90</td>
<td><strong>85.71</strong></td>
<td>377.27</td>
</tr>
</tbody>
</table>
Table 3. Mean forecasting performance metrics for the sub-epidemic models (ensemble weights were based on the relative likelihood) and the ARIMA models across 98 sequential weekly calibration periods of the daily time series of COVID-19 deaths in the USA from 20-April-2020 through 22-February 2022. Values highlighted in bold correspond to the best performance metrics.

**Discussion**

Our ensemble sub-epidemic modeling approach outperformed individual top-ranking sub-epidemic models and a set of ARIMA models in weekly short-term forecasts covering the national trajectory of the COVID-19 pandemic in the USA from the early growth phase up until
the Omicron-dominated wave. This framework has demonstrated reliable forecasting performance across different pandemic phases from the early growth phase characterized by exponential or sub-exponential growth dynamics to plateaus and new disease surges driven by the relaxation of social distancing policies or the emergence of new variants. Importantly, we found that forecasting performance consistently improved for the ensemble sub-epidemic models that incorporated a higher number of top-ranking sub-epidemic models. The ensemble model incorporating the top four ranking sub-epidemic models consistently yielded the best performance, particularly in terms of the coverage rate of the 95% prediction interval and the weighted interval score.

Our findings support the power of ensemble modeling approaches (e.g.,[14-17]). Our ensemble modeling framework derived from a family of sub-epidemic models demonstrated improved performance as the number of top-ranking sub-epidemic models included in the ensemble increased. Prior studies have documented the potential of ensemble models to enhance forecasting performance during multi-epidemic periods [14]. For instance, in the context of influenza, one study utilized “weighted density ensembles” for predicting timing and severity metrics and found that the performance of the ensemble model was comparable to that of the top individual model, albeit the ensemble’s forecasts were more stable across influenza seasons [17]. In the context of dengue in Puerto Rico, another study found that forecasts derived from Bayesian averaging ensembles outperformed a set of individual models [25]. Results from the US COVID-19 Forecasting Hub CDC were consistent with our findings in that a multimodel ensemble frequently outperformed the set of individual models.
We also evaluated short-term forecasting performance by a set of ARIMA models, as prior studies have underscored the value of ARIMA models in epidemic forecasting [56], by providing a relatively simple and transparent approach to forecasting. For instance, in the context of influenza-like-illness in the USA, a set of ARIMA models provided reasonably accurate short-term forecasts during the 2016/17 influenza season [57]. In another forecasting study during multiple seasons of influenza in the USA, an ARIMA model yielded similar short-term forecasting performance compared to other models based on the mechanistic SIR modeling framework [58]. ARIMA models have also been used for spatial prediction of the COVID-19 epidemic [59, 60]. Another study [61] showed that the ARIMA model is more effective than the Prophet time series model for forecasting COVID-19 prevalence. Finally, it is worth noting that the US COVID-19 Forecast Hub did not include an ARIMA model in its set of evaluated models [49]. Therefore, it is interesting to assess how ARIMA models perform in the context of the COVID-19 pandemic in the US.

Prior work has underscored the need to assess alternative ways of constructing ensembles from a set of individual models [14, 16]. We explored two ways of constructing the ensembles by relying on the AICc or the relative likelihood associated with the individual models. We found that the short-term forecasting performance achieved by the ensemble models was not significantly affected by the type of ensemble weights used to construct them although performance using ensemble weights based on the reciprocal of the AICc was slightly better. Further research could explore how different weighting strategies influence the forecasting performance of ensemble modeling approaches.
Short-term forecasting is an essential attribute of the models. As prior studies have underscored, longer-term forecasts are of value, but their dependability varies inversely with the time horizon. Our 20 and 30-day forecasts are most valuable for monitoring, managing, and informing the relaxing of social distancing requirements. The early detection of potential disease resurgence can signal the need for strict distancing controls, and the reports of cases can identify the geographic location of incubating sub-epidemics.

Our study is not exempt of limitations. Our analysis relied on daily time series data of COVID-19 deaths in the USA, which is inherently noisy due to heterogeneous data reporting at fine spatial scales (i.e., county-level) [62]. Noisy data complicate the ability of any mathematical model to identify meaningful signals about the impact of transmission dynamics and control interventions. To deal with the high noise levels in the data, we fitted the models to smoothed time series rather than the actual daily series, as described in the parameter estimation section. Other forecasting studies, including the US COVID-19 Forecasting Hub, have relied on weekly death counts to address this issue [49]. Beyond the COVID-19 pandemic, there is a need to establish benchmarks to systematically assess forecasting performance across a diverse catalog of mathematical models and epidemic datasets involving multiple infectious diseases, social contexts, and spatial scales.

While our analysis demonstrated the accuracy of our ensemble sub-epidemic modeling framework in forecasting the COVID-19 pandemic, the same framework could be readily used to forecast other epidemics irrespective of the type of disease and spatial scale involved. Beyond infectious diseases, this framework could also be used to forecast other biological and social
growth processes, such as the epidemics of lung injury associated with e-cigarette use or vaping and the viral spread of information through social media platforms.

In summary, our ensemble sub-epidemic models provided reliable short-term forecasts of the trajectory of the COVID-19 pandemic in the USA involving multiple waves and outcompeted a set of ARIMA models. The forecasting performance of the ensemble models improved with the number of top-ranking sub-epidemic models included in the ensemble. This framework could be readily applied to investigate the spread of epidemics and pandemics beyond COVID-19 and in a range of problems in nature and society that would benefit from short-term predictions.

Author contributions: Conceptualization, G.C.; methodology, G.C., J.H, R.L, validation, G.C., R.L; formal analysis, G.C., R.L; investigation, G.C., R.L; resources, G.C., ; data curation, G.C., S.D.; writing—original draft preparation, G.C., R.L; writing, review, and editing, A.T., G.C., S.D., K.R., R.L., J.M.H., ; visualization, G.C, R.L; supervision, G.C., R.L; project administration, G.C.; funding acquisition, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: G.C. is partially supported from NSF grants 1610429 and 1633381 and R01 GM 130900. A.T. and S.D. are supported by a 2CI fellowship from Georgia State University.

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

References


30


21. Smith T, Ross A, Maire N, Chitnis N, Studer A, Hardy D, et al. Ensemble modeling of the likely public health impact of a pre-erythrocytic malaria vaccine. PLoS Med. 2012;9(1):e1001157. Epub 2012/01/25. doi: 10.1371/journal.pmed.1001157. PubMed PMID: 22272189; PubMed Central PMCID: PMC3260300 employed by the PATH Malaria Organization which was supporting the development of RTS,S, the vaccine which is the focus of this paper. AB left PATH prior to any collaboration on this paper. All other authors have declared no competing interests. The views expressed are those of the authors.


