

For the Special Issue: “Modeling and Simulation of Infectious Diseases-Propagation, decontamination and mitigation”

Characterization of Partially Observed Epidemics through Bayesian Inference - Application to COVID-19

Cosmin Safta · Jaideep Ray · Khachik Sargsyan

Received: date / Accepted: date

Abstract We demonstrate a Bayesian method for the “real-time” characterization and forecasting of partially observed COVID-19 epidemic. Characterization is the estimation of infection spread parameters using daily counts of symptomatic patients. The method is designed to help guide medical resource allocation in the early epoch of the outbreak. The estimation problem is posed as one of Bayesian inference and solved using a Markov chain Monte Carlo technique. The data used in this study was sourced before the arrival of the second wave of infection in July 2020. The proposed modeling approach, when applied at the country level, generally provides accurate forecasts at the regional, state and country level. The epidemiological model detected the flattening of the curve in California, after public health measures were instituted. The method also detected different disease dynamics when applied to specific regions of New Mexico.

Keywords Markov Chain Monte Carlo · pseudo-marginal MCMC · Bayesian framework · COVID-19 · infection rate · incubation model

1 Introduction

In this paper, we formulate and describe a data-driven epidemiological model to forecast the short-term evolution of a partially-observed epidemic, with the aim of helping estimate and plan the deployment of medical resources and personnel. It also allows

C. Safta
Sandia National Labs
E-mail: csafta@sandia.gov

J. Ray
Sandia National Labs
E-mail: jairay@sandia.gov

K. Sargsyan
Sandia National Labs
E-mail: ksargsy@sandia.gov

us to infer the efficacy of medical countermeasures deployed to control the outbreak. It also allows us to forecast, over a short period, the stream of patients seeking medical care, and thus estimate the demand for medical resources. It is meant to be used in the early days of the outbreak, when data and information about the pathogen and its interaction with its host population is scarce. The model is simple and makes few demands on our epidemiological knowledge of the pathogen. The method is cast as one of Bayesian inference of the latent infection rate (number of people infected per day), conditioned on a time-series of daily new (confirmed) cases of patients exhibiting symptoms and seeking medical care. The model is demonstrated on the COVID-19 pandemic that swept through the US in spring 2020. The model generalizes across a range of host population sizes, and is demonstrated at the country-scale as well as for a sparsely populated desert region in Northwestern New Mexico, USA.

Developing a forecasting method that is applicable in the early epoch of a partially-observed outbreak poses some peculiar difficulties. The evolution of an outbreak depends on the characteristics of the pathogen and its interaction with patterns of life (i.e., population mixing) of the host population, both of which are ill-defined during the early epoch. These difficulties are further amplified if the pathogen is novel, and its myriad presentations in a human population is not fully known. In such a case, the various stages of the disease (e.g., prodrome, symptomatic etc.), and the residence times in each, are unknown. Further, the patterns of life are expected to change over time as the virulence of the pathogen becomes known and medical countermeasures are put in place. In addition, to be useful, the model must provide its forecasts and results in a timely fashion, despite imperfect knowledge about the efficacy of the countermeasures and the degree of adherence of the population to them. These requirements point towards a simple model that does not require much information or knowledge of the pathogen and its behavior to produce its outputs. In addition, it also strongly suggest an inferential approach conditioned on an easily obtained/observed marker of the progression of the outbreak (e.g., the time-series of daily new cases), even though the quality of the observations may leave much to be desired.

In keeping with these insights into the peculiarities of forecasting during the early epoch, we pose our method as one of Bayesian inference of a parametric model of the latent infection rate (which varies over time). This infection rate curve is convolved with the PDF (Probability Density Function) of the incubation period of the disease to produce an expression for the time-series of newly symptomatic cases, an observable that is widely reported as “daily new cases” by various data sources [2,5,6]. A Markov chain Monte Carlo (MCMC) method is used to construct a distribution for the parameters of the infection rate curve, even under an imperfect knowledge of the incubation period’s PDF. This uncertain infection rate curve, which reflects the lack of data and the imperfections of the epidemiological model, can be used to provide stochastic, short-term forecasts of the outbreak’s evolution. The reliance on the daily new cases, rather than the time-series of fatalities (which, arguably, has fewer uncertainties in it) is deliberate. Fatalities are delayed and thus are not a timely source of information. In addition, in order to model fatalities, the presentation and progress of the disease in an individual must be completely known, a luxury not available for a novel pathogen. Our approach is heavily influenced by a similar effort undertaken in the late 1980s to analyze and forecast the progress of AIDS in San Francisco [12],

with its reliance on simplicity and inference, though the formulation of our model is original, as is the statistical method used in the inference.

There have been many attempts at modeling the behavior of COVID-19, most of which have forecasting as their primary aim. Our ignorance of its behavior in the human population is evident in the choice of modeling techniques used for the purpose. Time-series methods such as ARIMA [9, 39] and logistic regression for cumulative time-series [28] have been used extensively, as have machine-learning methods using Long Short-Term Memory models [16, 17] and autoencoders [18]. These approaches do not require any disease models and focus solely on fitting the data, daily or cumulative, of new cases as reported. Ref. [45] contains a comprehensive summary of various machine-learning methods used to “curve-fit” COVID-19 data and produce forecasts. Approaches that attempt to embed disease dynamical models into their forecasting process have also been explored, usually via compartmental SEIR models or their extensions. Compartmental models represent the progress of the disease in an individual via a set of stages with exponentially distributed residence times, and predict the size of the population in each of the stages. These mechanistic models are fitted to data to infer the means of the exponential distributions, using MCMC [11] and Ensemble Kalman Filters (or modifications) [19, 20, 38]. Less common disease modeling techniques, such as agent-based simulations [26], modeling of infection and epidemiological processes as statistical ones [3] and the propagation of epidemics on a social network [43] have also been explored, as have been methods that include proxies of population mixing (e.g., using Google mobility data [15]). There is also a group of 20 modeling teams that submit their epidemiological forecasts regarding the COVID-19 pandemic to the CDC; their details can be found at their website [7].

Apart from forecasting and assisting in resource allocation, data-driven methods have also been used to assess whether countermeasures have been successful e.g., by replacing a time-series of daily new cases with a piecewise linear approximation, the author in Ref. [14] showed that the lockdown in India did not have a significant effect on “flattening the curve”. We perform a similar analysis later, for the shelter-in-place orders implemented in California in mid-March, 2020. Efforts to develop metrics, derived from observed time-series of cases, that could be used to monitor countermeasures’ efficacy and trigger decisions [13], too exist. There have also been studies to estimate the unrecorded cases of COVID-19 by computing excess cases of Influenza-Like-Illness versus previous years’ numbers [32]. Estimates of resurgence of the disease as nations come out of self-quarantine have also been developed [47].

Some modeling and forecasting efforts have played an important role in guiding policy makers when responding to the pandemic. The first COVID-19 forecasts, which led to serious considerations of imposing restrictions on the mixing of people in the United Kingdom and the USA, were generated by a team from Imperial College, London [21]. Influential COVID-19 forecasts for USA were generated by a team from the University of Washington, Seattle [36] and were used to estimate the demand for medical resources [35]. These forecasts have also been compared to actual data, once they became available [34], an assessment that we also perform in this paper. Adaptations of the University of Washington model, that include mobility data to assess changes in population mixing, have also been developed [46], show-

ing enduring interest in using models and data to understand, predict and control the pandemic.

Fig. 1 shows a schematic of the overall workflow developed in this paper. The epidemiological model is formulated in §2, with postulated forms for the infection rate curve and the derivation of the prediction for daily new cases; we also discuss a filtering approach that is applied to the data before using it to infer model parameters. In §3 we describe the “error model” and the statistical approach used to infer the latent infection rate curve, and to account for the uncertainties in the incubation period distribution. Results, including push-forward posteriors and posterior predictive checks,

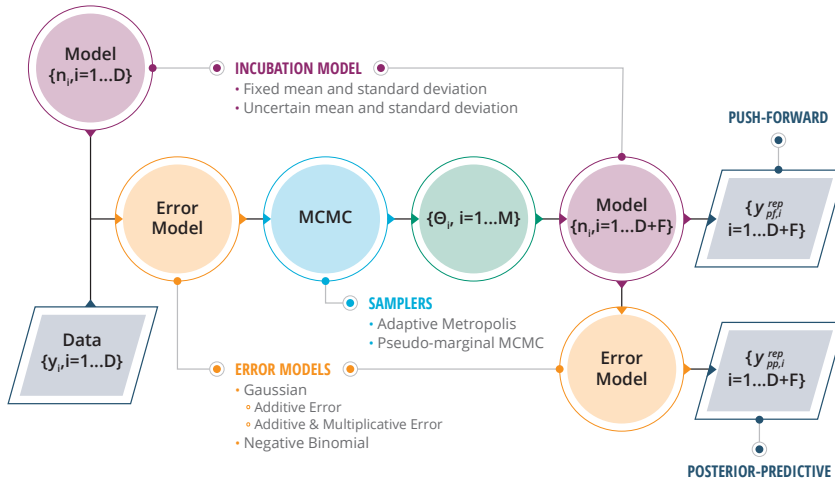


Fig. 1: Epidemiological model inference and forecast workflow. The “Model” circle encompasses the convolution between the infection rate and the incubation model described in §2 and the “Error Model” circle illustrates the choices for the discrepancy between the model and the data presented in §3.1.

are presented in §4 and we conclude in §5. The appendix includes a presentation of data sources used in this paper.

2 Modeling Approach

We present here an epidemiological model to characterize and forecast the rate at which people turn symptomatic from disease over time. For the purpose of this work, we assume that once people develop symptoms, they have ready access to medical services and can be diagnosed readily. From this perspective, these forecasts represent a lower bound on the actual number of people that are infected with COVID-19 as the people currently infected, but still incubating, are not accounted for. A fraction

of the population infected might also exhibit minor or no symptoms at all and might not seek medical advice. Therefore, these cases will not be part of patient counts released by health officials. The epidemiological model consists of two main components: an infection rate model, presented in §2.1 and an incubation rate model, described in §2.2. These models are combined, through a convolution presented in §2.3, into a forecast of number of cases that turn symptomatic daily. These forecasts are compared to data presented in §2.4 and the Appendix.

2.1 Infection Rate Model

The infection rate represents the probability of an individual, that eventually will get affected during an epidemic, to get infected at a specific time following the start of the epidemic [30]. We approximate the rate of infection with a gamma distribution with unknown shape parameter k and scale parameter θ . Depending on the choice for the pair (k, θ) this distribution can display a sharp increase in the number of people infected followed by a long tail, a dynamic that could lead to significant pressure on the medical resources. Alternatively, the model can also capture weaker gradients (“flattening the curve”) equivalent to public health efforts to temporarily increase social separation and thus reducing the pressure on available medical resources.

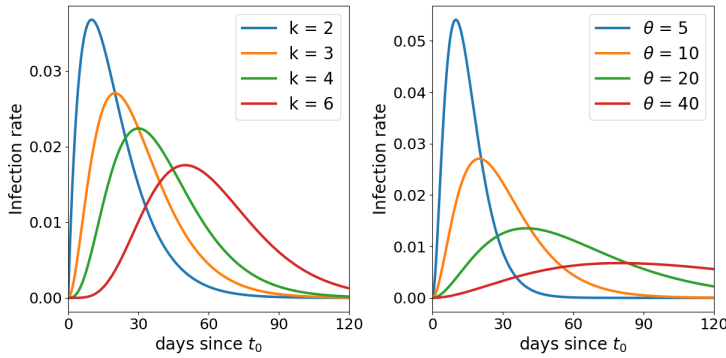


Fig. 2: Infection rate models with fixed scale parameters $\theta = 10$ (left frame) and fixed shape parameter $k = 3$ (right frame).

The infection rate model is given by

$$f_{\Gamma}(t; k, \theta) = \theta^{-k} t^{k-1} \exp(-t/\theta) / \Gamma(k) \quad (1)$$

where $f_{\Gamma}(t; k, \theta)$ is the probability density function (pdf) of the gamma distribution, with k and θ strictly positive. Fig. 2 shows example infection rate models for several shape and scale parameter values. The time in this figure is referenced with respect to start of the epidemic, t_0 . Larger k and θ values lead to mild gradients but extend the duration of the infection rate curves.

Table 1: Nominal and 95% Confidence Interval (CI) values (from Reference [29]) for the parameters of the incubation model (in Eq. (2)).

| Parameter | Nominal | 95% CI |
|-----------|---------|---------------|
| μ | 1.621 | (1.504,1.755) |
| σ | 0.418 | (0.271,0.542) |

2.2 Incubation Model

Most of the results presented in this paper employ a lognormal incubation distribution for COVID-19 [29]. The nominal and 95% Confidence Interval values for the mean, μ , and standard deviation σ of the natural logarithm of the incubation model are provided in Table 1.

The PDF, f_{LN} , and cumulative distribution function (CDF), F_{LN} , of the lognormal distribution are given by

$$f_{LN}(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right), F_{LN}(t; \mu, \sigma) = \frac{1}{2} \operatorname{erfc}\left(-\frac{\log t - \mu}{\sigma\sqrt{2}}\right) \quad (2)$$

To ascertain the impact of limited sample size on the uncertainty of μ and σ , we analyze their theoretical distributions and compare with the data in Table 1. Let $\hat{\mu}$ and $\hat{\sigma}$ be the mean and standard deviation computed from a set of n samples of the natural logarithm of the incubation rate random variable. It follows that

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$$

has a Student's t-distribution with n -degrees of freedom. To model the uncertainty in $\hat{\sigma}$ we assume that

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$$

has a χ^2 distribution with $(n-1)$ degrees of freedom. While the data in [29] is based on $n = 181$ confirmed cases, we found the corresponding 95% CIs for μ and σ computed based on the Student's t and chi-square distributions assumed above to be narrower than the ranges provided in Table 1. Instead, to construct uncertain models for these statistics, we employed a number of degrees of freedom $n^* = 36$ that provided the closest agreement, in a L_2 -sense, to the 95% CI in the reference. The resulting 95% CIs for μ and σ based on this fit are [1.48, 1.76] and [0.320, 0.515], respectively.

The left frame in Fig. 3 shows the family of PDFs with μ and σ drawn from Student's t and χ^2 distributions, respectively. The nominal incubation PDF is shown in black in this frame. The impact of the uncertainty in the incubation model parameters is displayed in the right frame of this figure. For example, 7 days after infection, there is a large variability (60%-90%) in the fraction of infected people that completed the incubation phase and started displaying symptoms. This variability decreases at later times, e.g. after 10 days more than 85% of case completed the incubation process.

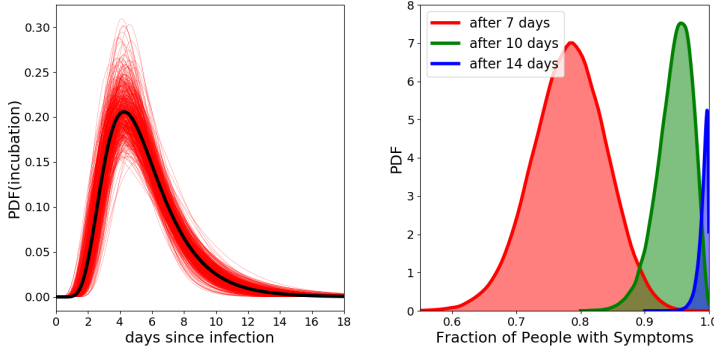


Fig. 3: Probability density functions for the incubation model (left frame) and fraction of people for which incubation ran its course after 7, 10, and 14 days respectively (right frame).

In the results section we will compare results obtained with the lognormal incubation model with results based on other probability distributions. Again, we turn to [29] which provides parameter values corresponding to gamma, Weibull, and Erlang distributions.

2.3 Daily Symptomatic Cases

With these assumptions the number of people infected *and* with completed incubation period at time t_i can be written as a convolution between the infection rate and the cumulative distribution function for the incubation distribution [12, 42, 44]

$$N_i = N \int_{t_0}^{t_i} f_{\Gamma}(\tau - t_0; k, \theta) F_{LN}(t_i - \tau; \mu, \sigma) d\tau \quad (3)$$

where N is the total number of people that will be infected throughout the epidemic and t_0 is the start time of the epidemic. This formulation assumes independence between the calendar date of the infection and the incubation distribution.

Using Eq. (3), the number of people developing symptoms between times t_{i-1} and t_i is computed as

$$n_i = N_i - N_{i-1} = N \int_{t_0}^{t_i} f_{\Gamma}(\tau - t_0; k, \theta) (F_{LN}^*(t_i - \tau; \mu, \sigma) - F_{LN}^*(t_{i-1} - \tau; \mu, \sigma)) d\tau \quad (4)$$

where

$$F_{LN}^*(t; \mu, \sigma) = \begin{cases} 0 & \text{if } t < 0 \\ F_{LN}(t; \mu, \sigma) & \text{if } t \geq 0 \end{cases} \quad (5)$$

In Eq. (4), the second term under the integral, $F_{LN}^*(t_i - \tau; \mu, \sigma) - F_{LN}^*(t_{i-1} - \tau; \mu, \sigma)$ can be approximated using the lognormal pdf as

$$F_{LN}^*(t_i - \tau; \mu, \sigma) - F_{LN}^*(t_{i-1} - \tau; \mu, \sigma) \approx (t_i - t_{i-1}) f_{LN}(t_i - \tau; \mu, \sigma) \quad (6)$$

leading to

$$n_i \approx N(t_i - t_{i-1}) \int_{t_0}^{t_i} f_{\Gamma}(\tau - t_0; k, \theta) f_{LN}(t_i - \tau; \mu, \sigma) d\tau \quad (7)$$

where f_{LN} is the lognormal pdf. The results presented in §4 compute the number of people that turn symptomatic daily, i.e. $t_i - t_{i-1} = 1$ [day].

2.4 Data

The number of people developing symptoms daily n_i , computed through Eqs. (4) or (7), are compared to data obtained from several sources at the national, state, or regional levels. We present the data sources in the Appendix.

We found that, for some states or regions, the reported daily counts exhibited a significant amount of noise. This is caused by variation in testing capabilities and sometimes by how data is aggregated from region to region and the time of the day when it is made available. Sometimes previously undiagnosed cases are categorized as COVID-19 and reported on the current day instead of being allocated to the original date. We employ a set of finite difference filters [25,41] that preserve low wavenumber information, i.e. weekly or monthly trends, and reduces high wavenumber noise, e.g. large day to day variability such as all cases for successive days being reported at the end of the time range

$$\hat{\mathbf{y}} = F \mathbf{y}, \quad F = \mathbb{I} + (-1)^{n+1} 2^{-2n} \mathbf{D} \quad (8)$$

Here \mathbf{y} is the original data, $\hat{\mathbf{y}}$ is the filtered data. Matrix \mathbb{I} is the identity matrix and \mathbf{D} is a band-diagonal matrix, e.g. triadiagonal for a $n=2$, i.e. a 2-nd order filter and pentadiagonal for a $n=4$, i.e. a 4-th order filter. We have compared 2-nd and 4-th order filters, and did not observe any significant difference between the filtered results. Reference [41] provides \mathbf{D} matrices for filters up to 12-th order.

Time series of \mathbf{y} and $\hat{\mathbf{y}}$ for several regions are presented in the Appendix. For the remainder of this paper we will only use filtered data to infer epidemiological parameters. For notational convenience, we will drop the hat and refer to the filtered data as \mathbf{y} .

Note that all the data used in this study predate June 1, 2020 (in fact, most of the studies use data gathered before May 15, 2020) when COVID-19 tests were administered primarily to symptomatic patients. Thus the results and inferences presented in this paper apply only to the symptomatic cohort who seek medical care, and thus pose the demand for medical resources. The data is also bereft of any information about the "second wave" of infections that affected Southern and Western USA in late June, 2020 [8].

3 Statistical Methodology

Given data, \mathbf{y} , in the form of time-series of daily counts, as shown in §2.4, and the model predictions n_i for the number of new symptomatic counts daily, presented in

§2, we will employ a Bayesian framework to calibrate the epidemiological model parameters. The discrepancy between the data and the model is written as

$$\mathbf{y} = \mathbf{n}(\Theta) + \varepsilon \quad (9)$$

where \mathbf{y} and \mathbf{n} are arrays containing the data and model predictions

$$\mathbf{y} = \{y(t_1), y(t_2), \dots, y(t_d)\}, \quad \mathbf{n} = \{n_1(\Theta), n_2(\Theta), \dots, n_d(\Theta)\}.$$

Here, d is the number of data points, the model parameters are grouped as $\Theta = \{t_0, N, k, \theta\}$ and ε represents the error model and encapsulates, in this context, both errors in the observations as well as errors due to imperfect modeling choices. The observation errors include variations due to testing capabilities as well as errors when tests are interpreted. Values for the vector of parameters Θ can be estimated in the form of a multivariate PDF via Bayes theorem

$$p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta)p(\Theta) \quad (10)$$

where $p(\Theta|\mathbf{y})$ is the posterior distribution we are seeking after observing the data \mathbf{y} , $p(\mathbf{y}|\Theta)$ is the likelihood of observing the data \mathbf{y} for a particular value of Θ , and $p(\Theta)$ encapsulates any prior information available for the model parameters. Bayesian methods are well-suited for dealing with heterogeneous sources of uncertainty, in this case from our modeling assumptions, i.e. model and parametric uncertainties, as well as the communicated daily counts of COVID-19 new cases, i.e. experimental errors.

3.1 Likelihood Construction

In this work we explore both deterministic and stochastic formulations for the incubation model. In the former case the mean and standard deviation of the incubation model are fixed at their nominal values and the model prediction n_i for day t_i is a scalar value that depends on Θ only. In the latter case, the incubation model is stochastic with mean and standard deviation of its natural logarithm treated as Student's t and χ^2 random variables, respectively, as discussed in §2.2. Let us denote the underlying independent random variables by $\xi = \{\xi_\mu, \xi_\sigma\}$. The model prediction $n_i(\xi)$ is now a random variable induced by ξ plugged in Eq. (4), and $\mathbf{n}(\xi)$ is a random vector.

We explore two formulations for the statistical discrepancy ε between \mathbf{n} and \mathbf{y} . In the first approach we assume ε has a zero-mean Multivariate Normal (MVN) distribution. Under this assumption the likelihood $p(\mathbf{y}|\Theta)$ for the *deterministic incubation model* can be written as

$$p(\mathbf{y}|\Theta) = \pi_{\mathbf{n}(\Theta)}(\mathbf{y}) = (2\pi)^{-D/2} |C_n|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{n}(\Theta))^T C_n^{-1} (\mathbf{y} - \mathbf{n}(\Theta))\right) \quad (11)$$

The covariance matrix C_n can in principle be parameterized, e.g. square exponential or Matern models, and the corresponding parameters inferred jointly with Θ . However, given the sparsity of data, we neglect correlations across time and presume a diagonal covariance matrix with diagonal entries computed as

$$C_{n,ii} = \sigma_i^2 = (\sigma_a + \sigma_m n_i(\Theta))^2 \quad (12)$$

The additive, σ_a , and multiplicative, σ_m , components will be inferred jointly with the model parameters Θ

$$\Theta = \{t_0, N, k, \theta\} \rightarrow \Theta = \{t_0, N, k, \theta, \log \sigma_a, \log \sigma_m\}$$

Here, we infer the logarithm of these parameters to ensure they remain positive. Under these assumptions, the MVN likelihood in Eq. (11) is written as a product of independent Gaussian densities

$$p(\mathbf{y}|\Theta) = \prod_{i=1}^D \pi_{n_i(\Theta)}(y_i) = (2\pi)^{-D/2} \prod_{i=1}^D \sigma_i^{-1} \exp\left(-\frac{(y_i - n_i)^2}{2\sigma_i^2}\right) \quad (13)$$

where σ_i is given by Eq. (12). In §4.3 we will compare results obtained using only the additive part σ_a , i.e. fixing $\sigma_m = 0$, of Eq. (12) with results using both the additive and multiplicative components.

The second approach assumes a negative-binomial distribution for the discrepancy between data and model predictions. The negative-binomial distribution is used commonly in epidemiology to model overly dispersed data, e.g. in case where the standard deviation exceeds the mean [31]. This is observed for most regions explored in this report, in particular for the second half of April and the first half of May. For this modeling choice, the likelihood of observing the data given a choice for the model parameters is given by

$$p(\mathbf{y}|\Theta) = \prod_{i=1}^D \pi_{n_i(\Theta)}(y_i) = \prod_{i=1}^D \binom{y_i + \alpha - 1}{\alpha - 1} \left(\frac{\alpha}{\alpha + n_i(\Theta)}\right)^\alpha \left(\frac{n_i(\Theta)}{\alpha + n_i(\Theta)}\right)^{y_i} \quad (14)$$

where $\alpha > 0$ is the dispersion parameter, and

$$\binom{y_i + \alpha - 1}{\alpha - 1} = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)\Gamma(y_i + 1)} \quad (15)$$

is the binomial coefficient. For simulations employing a negative binomial distribution of discrepancies, the logarithm of the dispersion parameter α (to ensure it remains positive) will be inferred jointly with the other model parameters, $\Theta = \{t_0, N, k, \theta, \log \alpha\}$.

For the *stochastic incubation model* the likelihood reads as

$$p(\mathbf{y}|\Theta) = \pi_{n(\Theta),\xi}(\mathbf{y}), \quad (16)$$

which we simplify by assuming independence of the discrepancies between different days, arriving at

$$\pi_{n(\Theta),\xi}(\mathbf{y}) = \prod_{i=1}^D \pi_{n_i(\Theta),\xi}(y_i). \quad (17)$$

Unlike the deterministic incubation model, the likelihood elements for each day $\pi_{n_i(\Theta),\xi}(y_i)$ are not analytically tractable anymore since they now incorporate contributions from ξ , i.e. from the variability of the parameters of the incubation model. One can evaluate the likelihood via kernel density estimation by sampling ξ for each sample of Θ , and combining these samples with samples of the assumed discrepancy ε , in order to arrive at an estimate of $\pi_{n_i(\Theta),\xi}(y_i)$. In fact, by sampling a *single* value of ξ for each sample of Θ , one achieves an unbiased estimate of the likelihood $\pi_{n_i(\Theta),\xi}(y_i)$, and given the independent-component assumption, it also leads to an unbiased estimate of the full likelihood $\pi_{n(\Theta),\xi}(\mathbf{y})$.

3.2 Posterior Sampling

A Markov Chain Monte Carlo (MCMC) algorithm is used to sample from the posterior density $p(\Theta|\mathbf{y})$. MCMC is a class of techniques that allows sampling from a posterior distribution by constructing a Markov Chain that has the posterior as its stationary distribution. In particular, we use a delayed-rejection adaptive Metropolis (DRAM) algorithm [23]. We have also explored additional algorithms, including transitional MCMC (TMCMC) [27,37] as well as ensemble samplers [22] that allow model evaluations to run in parallel as well as sampling multi-modal posterior distributions. As we revised the model implementation, the computational expense reduced by approximately two orders of magnitude, and all results presented in this report are based on posterior sampling via DRAM.

A key step in MCMC is the accept-reject mechanism via Metropolis-Hastings algorithm. Each sample of Θ , drawn from a proposal $q(\cdot|\Theta_i)$ is accepted with probability

$$\alpha(\Theta_{i+1}, \Theta_i) = \min \left(1, \frac{p(\Theta_{i+1}|\mathbf{y})q(\Theta_i|\Theta_{i+1})}{p(\Theta_i|\mathbf{y})q(\Theta_{i+1}|\Theta_i)} \right)$$

where $p(\Theta_i|\mathbf{y})$ and $p(\Theta_{i+1}|\mathbf{y})$ are the values of the posterior pdf's evaluated at samples Θ_i and Θ_{i+1} , respectively. In this work we employ symmetrical proposals, $q(\Theta_i|\Theta_{i+1}) = q(\Theta_{i+1}|\Theta_i)$. This is a straightforward application of MCMC for the deterministic incubation model. In stochastic incubation model, we employ the unbiased estimate of the approximate likelihood as described in the previous section. This is the essence of the pseudo-marginal MCMC algorithm [10] guaranteeing that the accepted MCMC samples correspond to the posterior distribution. In other words, at each MCMC step we draw a random sample ξ from its distribution, and then we estimate the likelihood in a way similar to the deterministic incubation model, in Eqs. (13) or (14).

Fig. 4 shows samples corresponding to a typical MCMC simulation to sample the posterior distribution of Θ . We used the Raftery-Lewis diagnostic [40] to determine the number of MCMC samples required for converged statistics corresponding to stationary posterior distributions for Θ . The required number of samples is of the order $o(10^5 - 10^6)$ depending on the geographical region employed in the inference. The resulting Effective Sample Size [24] varies between 8,000 and 15,000 samples depending on each parameter which is sufficient to estimate joint distributions for the model parameters. Fig. 4 displays 1D and 2D joint marginal distributions based on the

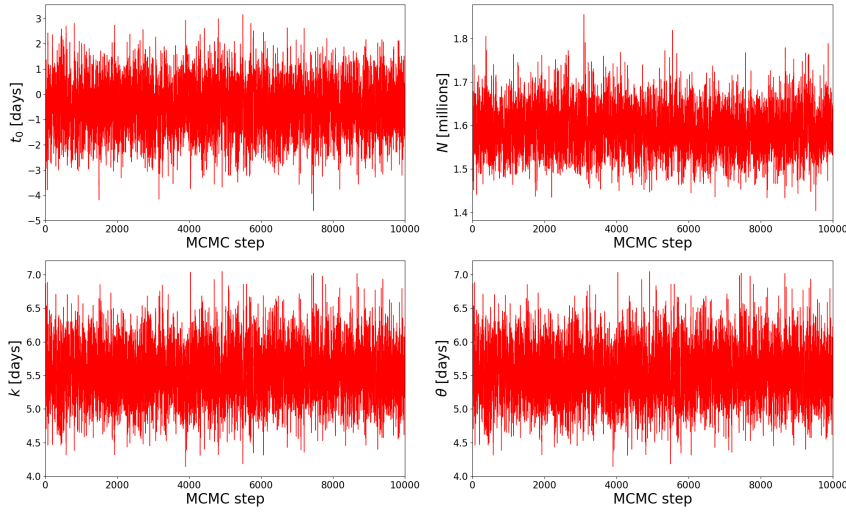


Fig. 4: MCMC samples for a simulation using US data up to May 1, 2020. The chain employed 10^6 samples; we skipped the 1st half and the remaining samples were thinned out to 10^4 samples. The t_0 samples are relative to March 1, 2020.

chain samples shown in the previous figure. These results indicate strong dependencies between some of the model parameters, e.g. between the start of the epidemic t_0 and the the scale parameter k of the infection rate model. This was somewhat expected based on the evolution of the daily counts of symptomatic cases and the functional form that couples the infection rate and incubation models. The number of samples in the MCMC simulations is tailored to capture these dependencies.

3.3 Predictive Assessment

We will employ both pushed-forward distributions and Bayesian posterior-predictive distributions [33] to assess the predictive skill of the proposed statistical model of the COVID-19 disease spread. The schematic in Eq. (18) illustrates the process to generate push-forward posterior estimates

$$p(\Theta|\mathbf{y}) \rightarrow \underbrace{\{\Theta^{(1)}, \dots, \Theta^{(m)}\}}_{\text{Posterior Samples}} \xrightarrow{n(\Theta)} \{\mathbf{y}^{(\text{pf},1)}, \dots, \mathbf{y}^{(\text{pf},m)}\} \rightarrow p_{\text{pf}}(\mathbf{y}^{(\text{pf})}|\mathbf{y}). \quad (18)$$

Here, $\mathbf{y}^{(\text{pf})}$ denotes hypothetical data \mathbf{y} and $p_{\text{pf}}(\mathbf{y}^{(\text{pf})}|\mathbf{y})$ denotes the push-forward probability density of the hypothetical data $\mathbf{y}^{(\text{pf})}$ conditioned on the observed data \mathbf{y} . We start with samples from the posterior distribution $p(\Theta|\mathbf{y})$. These samples are readily available from the MCMC exploration of the parameter space, i.e. similar to results shown in Fig. 4. Typically we subsample the MCMC chain to about 10-15K samples that will be used to generate push-forward statistics. Using these samples, we

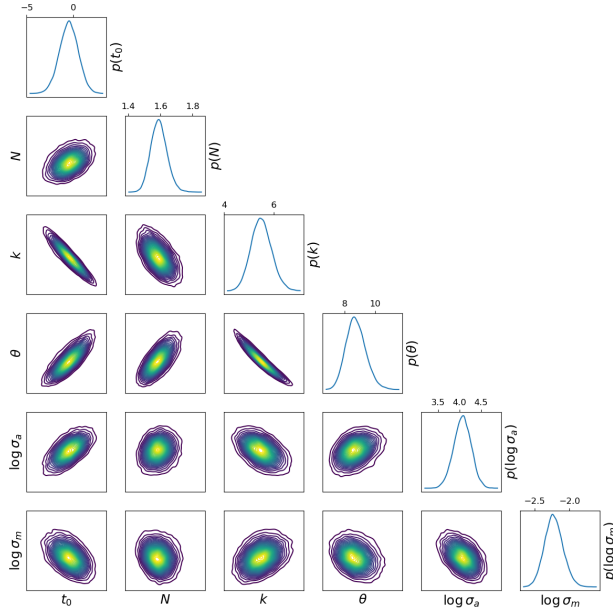


Fig. 5: 1D and 2D joint marginal distributions the components of $\Theta = \{t_0, N, k, \theta, \log \sigma_a, \log \sigma_m\}$.

evaluate the epidemiological model and collect the resulting $\mathbf{y}^{(\text{pf})} = \mathbf{n}(\Theta)$ samples that correspond to the push-forward posterior distribution $p_{\text{pf}}(\mathbf{y}^{(\text{pf})}|\mathbf{y})$.

The pushed-forward posterior does not account for the discrepancy between the data \mathbf{y} and the model predictions \mathbf{n} , subsumed into the definition of the error model ε presented in Eqs. (13) and (14). The Bayesian posterior-predictive distribution, defined in Eq. (19) is computed by marginalization of the likelihood over the posterior distribution of model parameters Θ :

$$p_{\text{pp}}(\mathbf{y}^{(\text{pp})}|\mathbf{y}) = \int_{\Theta} p(\mathbf{y}^{(\text{pp})}|\Theta)p(\Theta|\mathbf{y})d\Theta. \quad (19)$$

In practice, we estimate $p_{\text{pp}}(\mathbf{y}^{(\text{pp})}|\mathbf{y})$ through sampling, because analytical estimates are not usually available. The sampling workflow is similar to the one shown in Eq. (18). After the model evaluations $\mathbf{y} = \mathbf{n}(\Theta)$ are completed, we add random noise consistent with the likelihood model settings presented in §3.1. The resulting samples are used to compute summary statistics $p_{\text{pp}}(\mathbf{y}^{(\text{pp})}|\mathbf{y})$.

The push-forward and posterior-predictive distribution workflows can be used in hindcast mode, to check how well the model follows the data, and for short-term forecasts for the spread dynamics of this disease. In the hindcast regime, the infection rate is convolved with the incubation rate model to generate statistics for $\mathbf{y}^{(\text{pp})}$ (or $\mathbf{y}^{(\text{pf})}$) that will be compared against \mathbf{y} , the data used to infer the model parameters. The same functional form can be used to generate statistics for $\mathbf{y}^{(\text{pp})}$ (or $\mathbf{y}^{(\text{pf})}$) beyond

the set of dates for which data was available. We limit these forecasts to 7–10 days as our infection rate model does not count for changes in social dynamics that can significantly impact the epidemic over a longer time range.

4 Results

The statistical models described above are calibrated using data available at the country, state, and regional levels, and the calibrated model is used to gauge the agreement between the model and the data and to generate short-term forecasts, typically 7-10 days ahead.

First, we will assess the predictive capabilities of these models for several modeling choices:

- §4.2: Comparison of Incubation Models
- §4.3: Additive Error (AE) vs Additive+Multiplicative Error (A+ME) Models
- §4.4: Gaussian vs Negative Binomial Likelihood Models

We will then present results exploring the epidemiological dynamics at several geographical scales in §4.5.

4.1 Figure Annotations

The push-forward and posterior-predictive figures presented in this section show data used to calibrate the epidemiological model with filled black circles. The shaded color region illustrates either the pushed-forward posterior or the posterior-predictive distribution with darker colors near the median and lighter colors near the low and high quantile values. The blue colors correspond to the hindcast dates and red colors to forecasts. The inter-quartile range is marked with green lines and the 95% confidence interval with dashed lines. Some of the plots also show data collected at a later time, with open circles, to check the agreement between the forecast and the observed number of cases after the model has been calibrated.

4.2 Comparison of Incubation Models

We start the analysis with an assessment of the impact of the choice family of distributions on the model prediction. The left frame of Fig. 6 shows median (with red lines and symbols) and 95% CI with blue/magenta lines for the new daily cases based on lognormal, gamma, Weibull, and Erlang distributions for the incubation model. The mean and standard deviation of the natural logarithm of the associated lognormal random variable, and the shape and scale parameters for the other distributions are available in Appendix Table 2 from Reference [29]. The results for all four incubation models are visually very close. This observation holds for other simulations at national/state/regional levels (results not shown). The results presented in the remainder of this paper are based on lognormal incubation models. The right frame in Fig. 6

presents the corresponding infection rate curve that resulted from the model calibration. This represents a lower bound on the true number of infected people, as our model will not capture the asymptomatic cases or the population that displays minor symptoms and did not seek medical care.

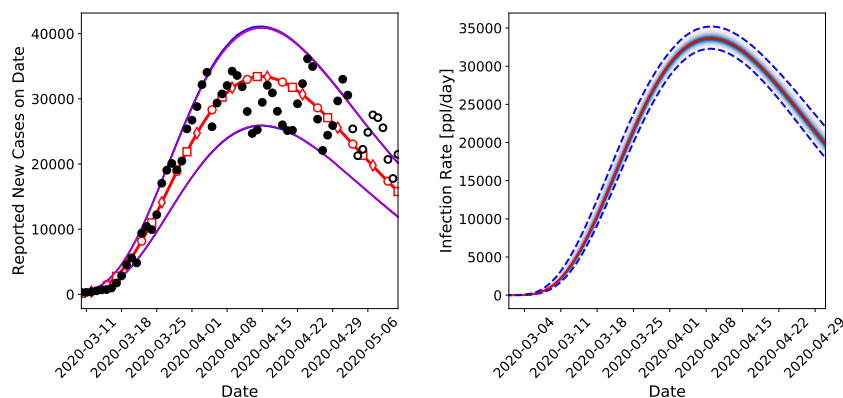


Fig. 6: (Left frame) Comparison of hindcasts/forecasts using several incubation models: the median is shown in red for lognormal (solid line) gamma (circle symbols), Weibull (square symbols), and Erlang (diamond symbols) distributions, respectively. The 95% CI is shown with blue lines for all lognormal and magenta for the other distributions; (Right frame) infection rate curve with calibrated shape and scale parameters shown in Fig. 5.

Next, we analyze the impact of the choice of deterministic vs stochastic incubation models on the model prediction. First we ran our model using the lognormal incubation model with mean and standard deviation fixed at their nominal values in Table 1. We then used the same dataset to calibrate the epidemiological model which employs an incubation rate with uncertain mean and standard deviation as described in §2.2. These results are labeled “Deterministic” and “Stochastic”, respectively, in Fig. 7. This figure shows results based on data corresponding to the United States. The choice of deterministic vs stochastic incubation models produce very similar outputs.

The results shown in the right frame of Fig. 3 indicate a relatively wide spread, between 0.64 and 0.95 with a nominal around 0.8, of the fraction of people that complete the incubation and start exhibiting symptoms 7 days after infection. Nevertheless, this variability does not have a significant impact on the model inference and subsequent forecasts. The noise induced by the stochastic incubation model is much smaller than the statistical noise introduced by the discrepancy between the data and the model. This observation holds for other datasets inspected for this work.

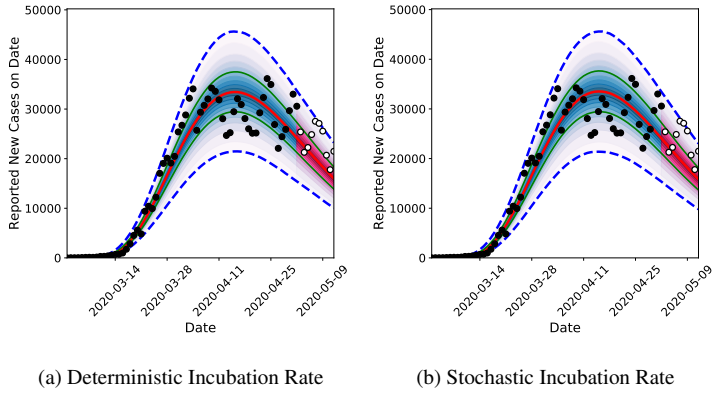


Fig. 7: Posterior-predictive forecast models using (a) nominal and (b) stochastic incubation rates. Epidemiological model inference employs aggregated data for the entire United States. Symbols and colors annotations are described in §4.1.

4.3 Additive vs Additive-Multiplicative Error Models

Next, we explore results based on either AE or A+ME formulations for the statistical discrepancy between the epidemiological model and the data. This choice impacts the construction of the covariance matrix for the Gaussian likelihood model, in Eq. (12). For AE we only infer σ_a while for A+ME we infer both σ_a and σ_m . The AE results in Fig. 8a are based on the same dataset as the A+ME results in Fig. 8b. Both

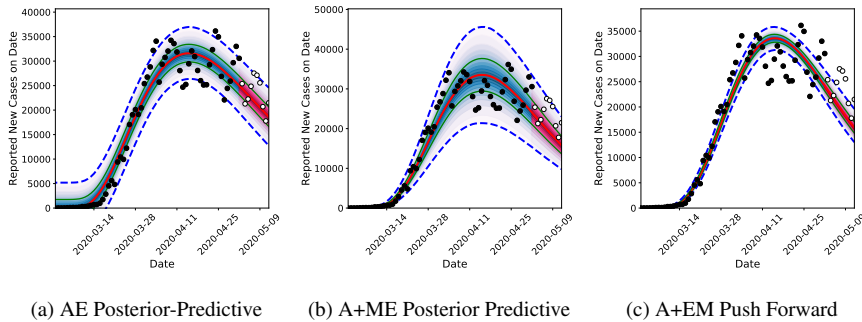


Fig. 8: Posterior-predictive and push-forward forecasts using data aggregated for all of the United States. The middle frame is the same as the right frame in Fig. 7 and is repeated here to facilitate the comparison of different modeling and forecast choices. Symbols and colors annotations are described in §4.1.

formulations present advantages and disadvantages when attempting to model daily symptomatic cases that span several orders of magnitude. The AE model, in Fig. 8a, presents a posterior-predictive range around the peak region that is consistent with

the spread in the data. However, the constant $\sigma = \sigma_a$ over the entire date range results in much wider uncertainties predicted by the model at the onset of the epidemic. The A+ME model handles the discrepancy better overall as the multiplicative error term allows it to adjust the uncertainty bound with the data. Nevertheless, this model results in a wider uncertainty band than warranted by the data near the peak region. These results indicate that a formulation for an error model that is time dependent can improve the discrepancy between the COVID-19 data and the epidemiological model.

We briefly explore the difference between pushed-forward posterior, in Fig. 8c, and the posterior-predictive data, in Fig. 7b. These results show that uncertainties in the model parameters alone are not sufficient to capture the spread in the data. This observation suggests more work is needed on augmenting the epidemiological model with embedded components that can explain the spread in the data without the need for external error terms.

4.4 Gaussian vs Negative Binomial Error Models

The negative binomial distribution is used commonly in epidemiology to model overly dispersed data, e.g. in cases where the variance exceeds the mean [31]. We also observe similar trends in some of the COVID-19 datasets. Fig. 9 shows results based on data for Alaska. The results based on the two error models are very similar, with the negative binomial results (on the top row) offering a slightly wider uncertainty band to better cover the data dispersion. Nevertheless, results are very similar, as they are for other regions that exhibit a similar number of daily cases, typically less than a few hundred. For regions with a larger number of daily cases, the likelihood evaluation was fraught with errors due to the evaluation of the negative binomial pdf. We therefore shifted our attention to the Gaussian formulation which offers a more robust evaluation for this problem.

4.5 Forecasts for Countries/States/Regions

In this section we examine forecasts based on data aggregated at country, state, and regional levels, and highlight similarities and differences in the epidemic dynamics resulted from these datasets.

4.5.1 Curve “flattening” in CA

The data in Fig. 10 illustrates the built-in delay in the disease dynamics due to the incubation process. A stay-at-home order was issued on March 19. Given the incubation rate distribution, it takes about 10 days for 90-95% of the people infected to start showing symptoms. After the stay at home order was issued, the number of daily case continued to rise because of infections that occurred before March 19. The data begins to flatten out in the first week of April and the model captures this trend a few days later, April 3-5. The data corresponding to April 9-11 show an increased

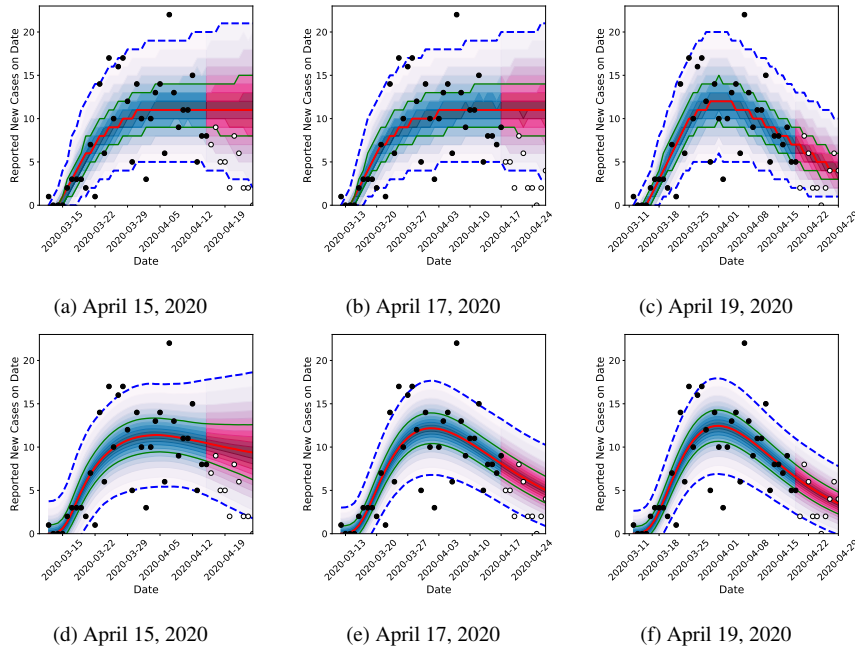


Fig. 9: Posterior-predictive forecasts for Alaska, using negative binomial likelihood (top row) and additive/multiplicative Gaussian likelihood (bottom row). Symbols and colors annotations are described in §4.1.

dispersion. To capture this increased noise, we switched from an AE model to A+ME model, with results shown in Fig. 11.

4.5.2 Example of Dynamics at Regional Scale: New Mexico

Figs. 12 and 13 present results showing the different dynamics of timing and scale of infections for the central (NM-C) and north-west (NM-NW) regions of New Mexico. These regions are also highlighted on the map in Fig. 20b. The data for the central region, shows a smaller daily count compared to the NW region. The epidemiological model captures the relatively large dispersion in the data for both regions. For the NM-C the first cases are recorded around March 10 and the model suggests the peak has been reached around mid-April, while NM-NW starts about a week later, around March 18, but records approximately twice more daily cases when it reaches the peak in the first half of May. Both regions display declining cases as of late May.

Comparing the Californian and New Mexican results, it is clear that the degree of scatter in the New Mexico data is much larger and adversely affects the inference, the model fit and the forecast accuracy. The reason for this scatter is unknown, but the daily numbers for New Mexico are much smaller than California's and are affected by individual events e.g., detection of transmission in a nursing home or a community. This is further accentuated by the fact that New Mexico is a sparsely populated re-

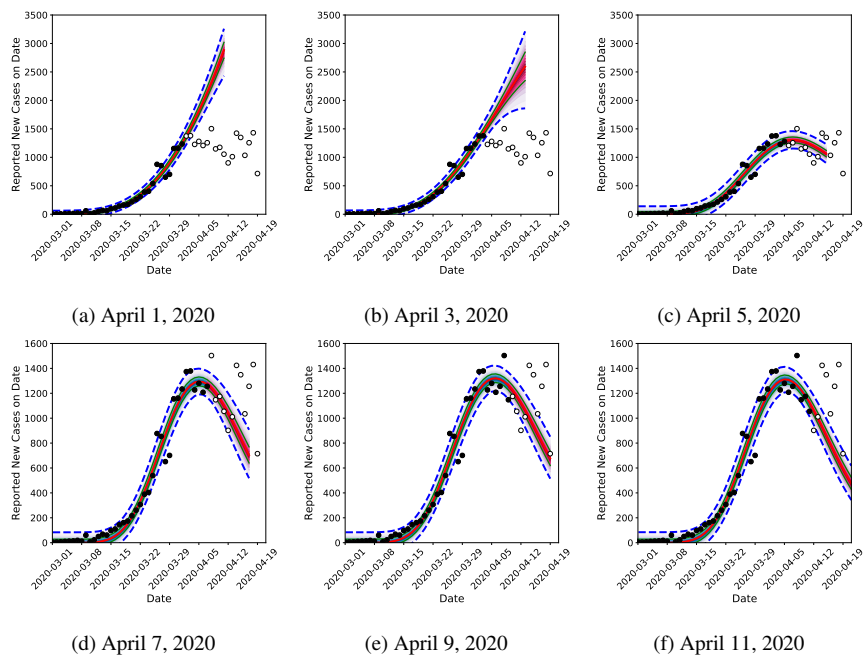


Fig. 10: Posterior-predictive forecasts for California, based on additive error models using data available on (a) April 1, 2020 through (f) April 11. Symbols and colors annotations are described in §4.1.

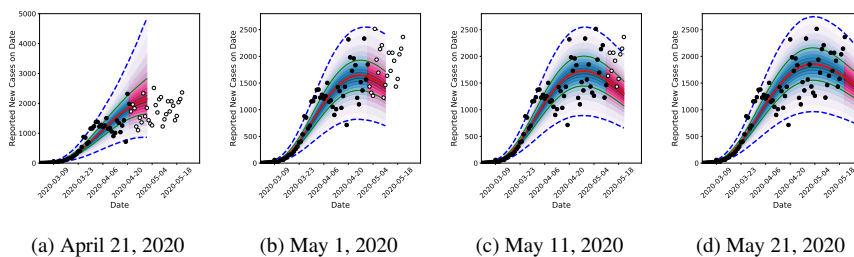


Fig. 11: Posterior-predictive forecasts for California, based on additive/multiplicative error models using data available on (a) April 21, 2020 through (d) May 21, 2020. Symbols and colors annotations are described in §4.1.

gion where sustained transmission, resulting in smooth curves, is largely impossible outside its few urban communities.

4.5.3 Moving target for US

This section discusses an analysis of the aggregate data from all US states. The posterior-predictive results shown in Fig. 14a—14d suggest the peak in the number

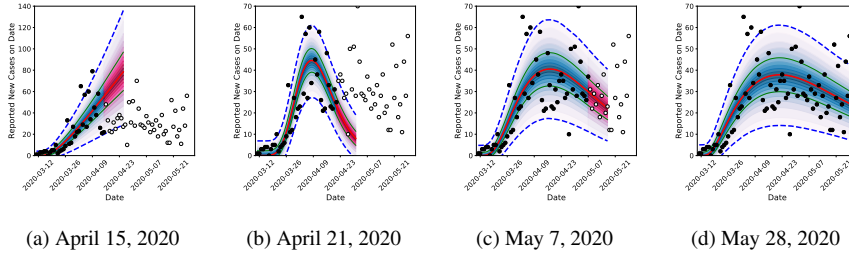


Fig. 12: Posterior-predictive forecasts for New Mexico central region, corresponding to counties highlighted with blue in Fig. 20b. Symbols and colors annotations are described in §4.1.

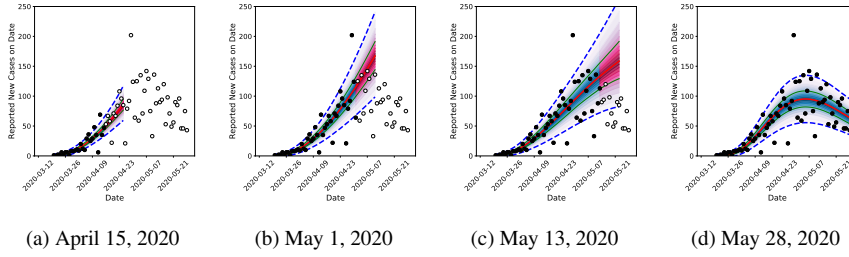


Fig. 13: Posterior-predictive forecasts for New Mexico north-west region, corresponding to counties highlighted with red in Fig. 20b. Symbols and colors annotations are described in §4.1.

of daily cases was reached around mid-April. Nevertheless the model had to adjust the downward slope as the number of daily cases has been declining at a slower pace compared to the time window that immediately followed the peak. As a result, the prediction for the total number of people, N , that would be infected in US during this first wave of infections has been steadily increasing as results show in Fig. 14e.

4.5.4 Sequence of Forecasts for Other Countries

We conclude our analysis of the proposed epidemiological model with available daily symptomatic cases pertaining to Germany, Italy, and Spain, in Figs. 15—17. For Germany, the uncertainty range increases while the epidemic is winding down, as the data has a relatively large spread compared to the number of daily cases recorded around mid-May. This reinforces an earlier point about the need to refine the error model with a time-dependent component. For Spain, a brief initial downslope can be observed in early April, also evident in the filtered data presented in Fig. 19b. This, however, was followed by large variations in the number of cases in the second half of April. This change could have been caused either by a scale-up of testing or by the occurrence of other infection hotspots in this country. This resulted in an overly-dispersed dataset and a wide uncertainty band for Spain. Forecasts based on daily symptomatic cases reported for Italy, in Fig. 17, exhibit an upward shift observed around April 10-20,

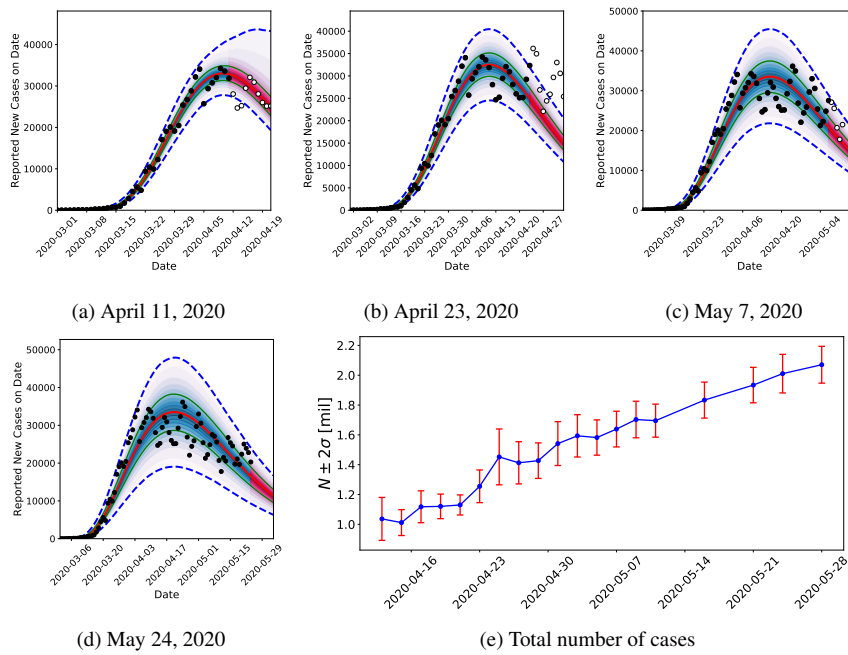


Fig. 14: (a-d) Posterior-predictive forecasts for US, based on additive/multiplicative error models and (e) Total number of cases N . Symbols and colors annotations for (a)-(d) are described in §4.1.

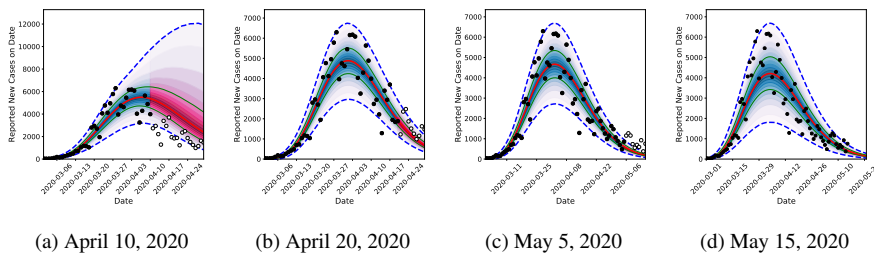


Fig. 15: Posterior-predictive forecasts for Germany, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1

similar to data for Spain above. The subsequent forecasts display narrower uncertainty bands compared to other similar forecasts above, possibly due to the absence of hotspots and/or regular data reporting.

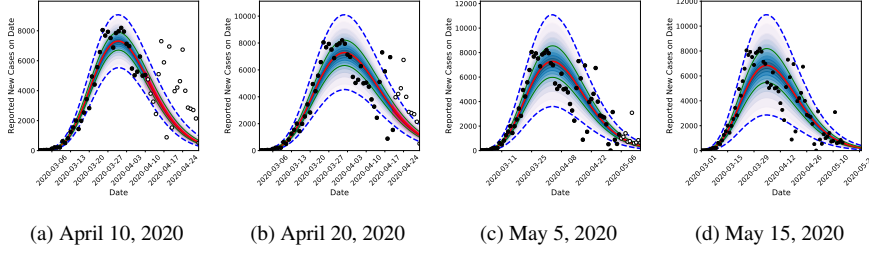


Fig. 16: Posterior-predictive forecasts for Spain, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1

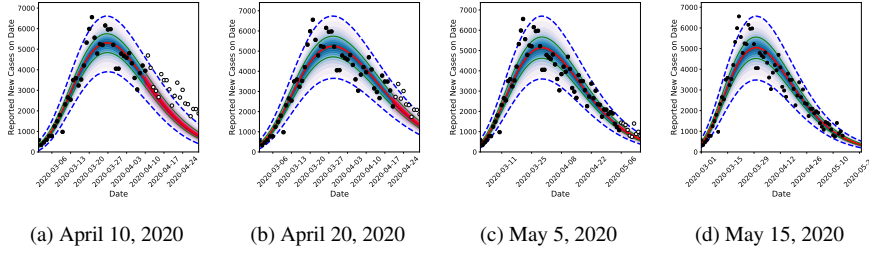


Fig. 17: Posterior-predictive forecasts for Italy, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1

4.6 Discussion

Fig. 10, 11 and 14 show inferences and forecasts obtained using data available till mid-May, 2020. They indicate that the outbreak was dying down, with forecasts of daily new cases trending down. In early June, public health measures to restrict population mixing were curtailed, and by mid-July, both California and the US were experiencing an explosive increase in new cases of COVID-19 being detected every day, quite at variance with the forecasts in the figures. This was due to the second wave of infections caused by enhanced population mixing.

The model in Eq. 3 cannot capture the second wave of infections due to its reliance on a unimodal infection curve $N f_{\Gamma}(\tau - t_0; k, \theta)$. This was by design, as the model is meant to be used early in an outbreak, with public health officials focussing on the first wave of infections. However, it can be trivially extended with a second infection curve to yield an augmented equation

$$N_i = N^{[1]} \int_{t_0}^{t_i} f_{\Gamma}^{[1]}(\tau - t_0; k^{[1]}, \theta^{[1]}) F_{LN}(t_i - \tau; \mu, \sigma) d\tau + N^{[2]} \int_{t_0}^{t_i} f_{\Gamma}^{[2]}(\tau - (t_0 + \Delta t); k^{[2]}, \theta^{[2]}) F_{LN}(t_i - \tau; \mu, \sigma) d\tau, \quad (20)$$

with two sets of parameters for the two infection curves, which are separated in time by $\Delta t > 0$. Eq. 20 is then fitted to data which is suspected to contain effects of two

waves of infection. This process does double the number of parameters to be estimated from data. However, the posterior density inferred for the parameters of the first wave (i.e., those with [1] superscript), using data collected before the arrival of the second wave, can be used to impose informative priors, considerably simplifying the estimation problem. Note that the augmentation shown in Eq. 20 is very intuitive, and can be repeated if multiple infection waves are suspected.

A second method that could, in principle, be used to infer multiple waves of infection are compartmental models e.g., SIR models, or their extensions. These models represent the epidemiological evolution of a patient through a sequence of compartments/states, with the residence time in each compartment modeled as a random variable. One of these compartments, “Infectious”, can then be used to model spread of the disease to other individuals. Such compartmental models have also been parlayed into Ordinary Differential Equation (ODE) models for an entire population, with the population distributed among the various compartments. ODE models assume that the residence time in each compartment is exponentially distributed, and using multiple compartments, can represent incubation and symptomatic periods that are not exponentially distributed. This does lead to an explosion of compartments. The spread-of-infection model often involves a time-dependent reproductive number $R(t)$ that can be used to model the effectiveness of epidemic control measures. It denotes that number of individuals a single infected individual will spread the disease to, and as public health measures are put in place (or removed), $R(t)$ will decrease or increase.

We did not consider SIR models, or their extensions, in our study as our model is meant to be used early in an outbreak when data is scarce and incomplete. Since our method is data-driven and involves fitting a model, a deterministic (ODE) compartmental model with few parameters would be desirable. The reasons for avoiding ODE-based compartmental models are:

- The incubation period of COVID-19 is *not* exponential (it is lognormal) and there is no way of modeling it with a single “Infectious” compartment.
- While it is possible to decompose the “Infectious” compartment into multiple sub-compartments, it would increase the dimensionality of the inverse problem as we would have to infer the fraction of the infected population in each of the sub-compartments. This is not desirable when data is scarce.
- We did not consider using extensions of SIR i.e., those with more compartments since it would require us to know the residence time in each compartment. This information is not available with much certainty at the start of the epidemic. This is particularly true for COVID-19 where only a small fraction of the “Infectious” cohort progress to compartments which exhibit symptoms.
- SIR models can infer the existence of a second wave of infections but would require a very flexible parameterization of $R(t)$ that would allow bi- or multimodal behavior. It is unknown what sparsely parameterized functional form would be sufficient for modeling $R(t)$.

5 Summary

This paper illustrates the performance of a method for producing short-term forecasts (with a forecasting horizon of about 7-10 days) of a partially-observed infectious disease outbreak. We have applied the method to the COVID-19 pandemic of spring, 2020. The forecasting problem is formulated as a Bayesian inverse problem, predicated on an incubation period model. The Bayesian inverse problem is solved using Markov chain Monte Carlo and infers parameters of the latent infection-rate curve from an observed time-series of new case counts. The forecast is merely the posterior-predictive simulations using realizations of the infection-rate curve and the incubation period model. The method accommodates multiple, competing incubation period models using a pseudo-marginal Metropolis-Hastings sampler. The variability in the incubation rate model has little impact on the forecast uncertainty, which is mostly due to the variability in the observed data and the discrepancy between the latent infection rate model and the spread dynamics at several geographical scales. The uncertainty in the incubation period distribution also has little impact on the inferred latent infection rate curve.

The method is applied at the country, provincial and regional/county scales. The bulk of the study used data aggregated at the state and country level for the United States, as well as counties in New Mexico and California. We also analyzed data from a few European countries. The wide disparity of daily new cases motivated us to study two formulations for the error model used in the likelihood, though the Gaussian error models was found to be acceptable for all cases. The most successful error model included a combination of multiplicative and additive errors. This was because of the wide disparity in the daily case counts experienced over the full duration of the outbreak. The method was found to be sufficiently robust to produce useful forecasts at all three spatial resolutions, though high-variance noise in low-count data (poorly reported / low-count / largely unscathed counties) posed the stiffest challenge in discerning the latent infection rate.

The method produces rough-and-ready information required to monitor the efficacy of quarantining efforts. It can be used to predict the potential shift in demand of medical resources due to the model's inferential capabilities to detect changes in disease dynamics through short-term forecasts. It took about 10 days of data (about the 90% quantile of the incubation model distribution) to infer the flattening of the infection rate in California after curbs on population mixing were instituted. The method also detected the anomalous dynamics of COVID-19 in northwestern New Mexico, where the outbreak has displayed a stubborn persistence over time.

Our approach suffers from two shortcomings. The first is our reliance on the time-series of daily new confirmed cases as the calibration data. As the pandemic has progressed and testing for COVID-19 infection has become widespread in the USA, the daily confirmed new cases are no longer mainly of symptomatic cases who might require medical care, and forecasts developed using our method would overpredict the demand for medical resources. However, as stated in §1, our approach, with its emphasis on simplicity and reliance on easily observed data, is meant to be used in the early epoch of the outbreak for medical resource forecasting, and within those pragmatic considerations, has worked well. The approach could perhaps be augmented

with a time-series of COVID-19 tests administered every day to tease apart the effect of increased testing on the observed data, but that is beyond the scope of the current work. Undoubtedly this would result in a more complex model, which would need to be conditioned on more plentiful data, which might not be readily available during the early epoch of an outbreak.

The second shortcoming of our approach is that it does not model, detect or infer a second wave of infections, caused by an increase in population mixing. This can be accomplished by adding a second infection rate curve/model to the inference procedure. This doubles the number of parameters to be inferred from the data, but the parameters of the first wave can be tightly constrained using informative priors. This issue is currently being investigated by the authors.

Acknowledgements The authors acknowledge the helpful feedback that John Jakeman have provided on various aspects related to speed up of model evaluations. The authors also acknowledge the support Erin Acquesta, Thomas Catanach, Kenny Chowdhary, Bert Debusschere, Edgar Galvan, Gianluca Geraci, Mohammad Khalil, and Teresa Portone provided in scaling up the short-term forecasts to large datasets. This work was funded in part by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Khachik Sargsyan was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program through the FASTMath Institute. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. 2019-20 coronavirus pandemic. https://en.wikipedia.org/wiki/2019-20_coronavirus_pandemic. Accessed: 2020-05-10
2. Coronavirus (Covid-19) Data in the United States. <https://github.com/nytimes/covid-19-data>. Accessed: 2020-05-10
3. Covid-19 confirmed and forecasted case data. URL <https://covid-19.bsvgateway.org>. Accessed July 1, 2020
4. COVID-19 Coronavirus Pandemic. <https://www.worldometers.info/coronavirus>. Accessed: 2020-05-10
5. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. <https://github.com/CSSEGISandData/COVID-19>. Accessed: 2020-05-10
6. Covid-19 pandemic data/united states medical cases. https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/United_States_medical_cases. Accessed: 2020-05-10
7. Forecasts of total deaths. URL <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html>. Accessed July 1, 2020
8. Reopenings stall as u.s. records nearly 50,000 cases of covid-19 in single day. URL <https://www.reuters.com/article/us-health-coronavirus-usa/reopenings-stall-as-u-s-records-nearly-50000-cases-of-covid-19-in-single-day-idUSKBN2426LN>. Accessed July 1, 2020
9. Ajadi, N.A., Ogunsola, I.A., Damisa, S.A.: Modelling the occurrence of the novel pandemic covid-19 outbreak; a box and jenkins approach. medRxiv (2020). DOI 10.1101/2020.06.15.20131136. URL <https://www.medrxiv.org/content/early/2020/06/16/2020.06.15.20131136>
10. Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**(2), 697–725 (2009). DOI 10.1214/07-AOS574

11. Annan, J.D., Hargreaves, J.C.: Model calibration, nowcasting, and operational prediction of the covid-19 pandemic. medRxiv (2020). DOI 10.1101/2020.04.14.20065227. URL <https://www.medrxiv.org/content/early/2020/05/27/2020.04.14.20065227>
12. Brookmeyer, R., Gail, M.H.: A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the AIDS Epidemic. *Journal of the American Statistical Association* **83**(402), 301–308 (1988). DOI 10.1080/01621459.1988.10478599
13. Chang, S.R.: Development and application of pandemic projection measures (ppm) for forecasting the covid-19 outbreak. medRxiv (2020). DOI 10.1101/2020.05.30.20118158. URL <https://www.medrxiv.org/content/early/2020/06/03/2020.05.30.20118158>
14. Chaurasia, A.R.: Covid-19 trend and forecast in india: A joinpoint regression analysis. medRxiv (2020). DOI 10.1101/2020.05.26.20113399. URL <https://www.medrxiv.org/content/early/2020/06/03/2020.05.26.20113399>
15. Chiang, W.H., Liu, X., Mohler, G.: Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. medRxiv (2020). DOI 10.1101/2020.06.06.20124149. URL <https://www.medrxiv.org/content/early/2020/06/08/2020.06.06.20124149>
16. Deng, Q.: Dynamics and development of the covid-19 epidemics in the us: a compartmental model with deep learning enhancement. medRxiv (2020). DOI 10.1101/2020.05.31.20118414. URL <https://www.medrxiv.org/content/early/2020/06/06/2020.05.31.20118414>
17. Direkoglou, C., Sah, M.: Worldwide and regional forecasting of coronavirus (covid-19) spread using a deep learning model. medRxiv (2020). DOI 10.1101/2020.05.23.20111039. URL <https://www.medrxiv.org/content/early/2020/05/26/2020.05.23.20111039>
18. Distante, C., Gadelha Pereira, I., Garcia Goncalves, L.M., Piscitelli, P., Miani, A.: Forecasting covid-19 outbreak progression in italian regions: A model based on neural network training from chinese data. medRxiv (2020). DOI 10.1101/2020.04.09.20059055. URL <https://www.medrxiv.org/content/early/2020/04/14/2020.04.09.20059055>
19. Engbert, R., Rabe, M.M., Kliegl, R., Reich, S.: Sequential data assimilation of the stochastic seir epidemic model for regional covid-19 dynamics. medRxiv (2020). DOI 10.1101/2020.04.13.20063768. URL <https://www.medrxiv.org/content/early/2020/04/20/2020.04.13.20063768>
20. Evensen, G., Amezcu, J., Bocquet, M., Carrassi, A., Farchi, A., Fowler, A., Houtekamer, P., Jones, C.K.R.T., de Moraes, R., Pulido, M., Sampson, C., Vossepoel, F.: An international assessment of the covid-19 pandemic using ensemble data assimilation. medRxiv (2020). DOI 10.1101/2020.06.11.20128777. URL <https://www.medrxiv.org/content/early/2020/06/12/2020.06.11.20128777>
21. Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., et al.: Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. Tech. rep., Imperial College, London (2020). DOI 10.25561/77482. URL <http://hdl.handle.net/10044/1/77482>
22. Goodman, J., Weare, J.: Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science* **5**(1), 65–80 (2010). DOI 10.2140/camcos.2010.5.65
23. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242 (2001). DOI 10.2307/3318737
24. Kass, R., Carlin, B., Gelman, A., Neal, R.: Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician* **52**(2), 93–100 (1998). DOI 10.1080/00031305.1998.10480547
25. Kennedy, C.A., Carpenter, M.H.: Several new numerical methods for compressible shear-layer simulations. *Applied Numerical Mathematics* **14**(4), 397 – 433 (1994). DOI 10.1016/0168-9274(94)00004-2
26. Kerr, C.C., Stuart, R.M., Mistry, D., Abey Suriya, R.G., Hart, G., Rosenfeld, K., Selvaraj, P., Nunez, R.C., Hagedorn, B., George, L., Izzo, A., Palmer, A., Delpont, D., Bennette, C., Wagner, B., Chang, S., Cohen, J.A., Panovska-Griffiths, J., Jastrzebski, M., Oron, A.P., Wenger, E., Famulare, M., Klein, D.J.: Covasim: an agent-based model of covid-19 dynamics and interventions. medRxiv (2020). DOI 10.1101/2020.05.10.20097469. URL <https://www.medrxiv.org/content/early/2020/05/15/2020.05.10.20097469>
27. Khalil, M., Lao, J., Safta, C., Najm, H.: Transitional Markov Chain Monte Carlo Sampler in UQTK. Tech. Rep. SAND2020-3166, Sandia National Laboratories (2020)
28. Kriston, L.: Predictive accuracy of a hierarchical logistic model of cumulative sars-cov-2 case growth. medRxiv (2020). DOI 10.1101/2020.06.15.20130989. URL <https://www.medrxiv.org/content/early/2020/06/16/2020.06.15.20130989>

29. Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., Lessler, J.: The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* (2020). DOI 10.7326/M20-0504
30. Lloyd, A.L.: Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics. *Theoretical Population Biology* **60**(1), 59 – 71 (2001). DOI 10.1006/tpbi.2001.1525
31. Lloyd-Smith, J.O.: Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. *PLOS ONE* **2**(2), 1–8 (2007). DOI 10.1371/journal.pone.0000180
32. Lu, F.S., Nguyen, A.T., Link, N.B., Lipsitch, M., Santillana, M.: Estimating the early outbreak cumulative incidence of covid-19 in the united states: Three complementary approaches. *medRxiv* (2020). DOI 10.1101/2020.04.18.20070821. URL <https://www.medrxiv.org/content/early/2020/06/18/2020.04.18.20070821>
33. Lynch, S.M., Western, B.: Bayesian posterior predictive checks for complex models. *Sociological Methods and Research* **32**(3), 301–335 (2004). DOI 10.1177/0049124103257303
34. Marchant, R., Samia, N.I., Rosen, O., Tanner, M.A., Cripps, S.: Learning as we go: An examination of the statistical accuracy of covid19 daily death count predictions. *medRxiv* (2020). DOI 10.1101/2020.04.11.20062257. URL <https://www.medrxiv.org/content/early/2020/04/17/2020.04.11.20062257>
35. Murray, C.J., et al.: Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv* (2020). DOI 10.1101/2020.03.27.20043752. URL <https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752>
36. Murray, C.J., et al.: Forecasting the impact of the first wave of the covid-19 pandemic on hospital demand and deaths for the usa and european economic area countries. *medRxiv* (2020). DOI 10.1101/2020.04.21.20074732. URL <https://www.medrxiv.org/content/early/2020/04/26/2020.04.21.20074732>
37. Muto, M., Beck, J.L.: Bayesian Updating and Model Class Selection for Hysteretic Structural Models Using Stochastic Simulation. *Journal of Vibration and Control* **14**(1-2), 7–34 (2008). DOI 10.1177/1077546307079400
38. Pei, S., Shaman, J.: Initial simulation of sars-cov2 spread and intervention effects in the continental us. *medRxiv* (2020). DOI 10.1101/2020.03.21.20040303. URL <https://www.medrxiv.org/content/early/2020/03/27/2020.03.21.20040303>
39. Perone, G.: An arima model to forecast the spread and the final size of covid-2019 epidemic in italy. *medRxiv* (2020). DOI 10.1101/2020.04.27.20081539. URL <https://www.medrxiv.org/content/early/2020/05/03/2020.04.27.20081539>
40. Raftery, A., Lewis, S.: How many iterations in the gibbs sampler? In: *Bayesian Statistics 4*, pp. 763–773. Oxford University Press (1992)
41. Ray, J., Kennedy, C.A., Lefantzi, S., Najm, H.N.: Using High-Order Methods on Adaptively Refined Block-Structured Meshes: Derivatives, Interpolations, and Filters. *SIAM Journal on Scientific Computing* **29**(1), 139–181 (2007). DOI 10.1137/050647256
42. Ray, J., Lefantzi, S.: Deriving a model for influenza epidemics from historical data. Tech. Rep. SAND2011-6633, Sandia National Laboratories (2011)
43. Reich, O., Shalev, G., Kalvari, T.: Modeling covid-19 on a network: super-spreaders, testing and containment. *medRxiv* (2020). DOI 10.1101/2020.04.30.20081828. URL <https://www.medrxiv.org/content/early/2020/05/05/2020.04.30.20081828>
44. Safta, C., Ray, J., Sargsyan, K., Lefantzi, S., Cheng, K., Crary, D.: Real-time Characterization of Partially Observed Epidemics using Surrogate Models. Tech. Rep. SAND2011-6776, Sandia National Laboratories (2011)
45. Suzuki, Y., Suzuki, A.: Machine learning model estimating number of covid-19 infection cases over coming 24 days in every province of south korea (xgboost and multioutputregressor). *medRxiv* (2020). DOI 10.1101/2020.05.10.20097527. URL <https://www.medrxiv.org/content/early/2020/05/14/2020.05.10.20097527>
46. Woody, S., Garcia Tec, M., Dahan, M., Gaither, K., Lachmann, M., Fox, S., Meyers, L.A., Scott, J.G.: Projections for first-wave covid-19 deaths across the us using social-distancing measures derived from mobile phones. *medRxiv* (2020). DOI 10.1101/2020.04.16.20068163. URL <https://www.medrxiv.org/content/early/2020/04/26/2020.04.16.20068163>

47. Yamana, T., Pei, S., Kandula, S., Shaman, J.: Projection of covid-19 cases and deaths in the us as individual states re-open may 4,2020. medRxiv (2020). DOI 10.1101/2020.05.04.20090670. URL <https://www.medrxiv.org/content/early/2020/05/13/2020.05.04.20090670>

Appendix: Epidemiological Data

We have used several sources [1,2,4–6] to gather daily counts of symptomatic cases at several times while we performed this work. The illustrations in this section present both the original data with blue symbols as well as filtered data with red symbols and lines.

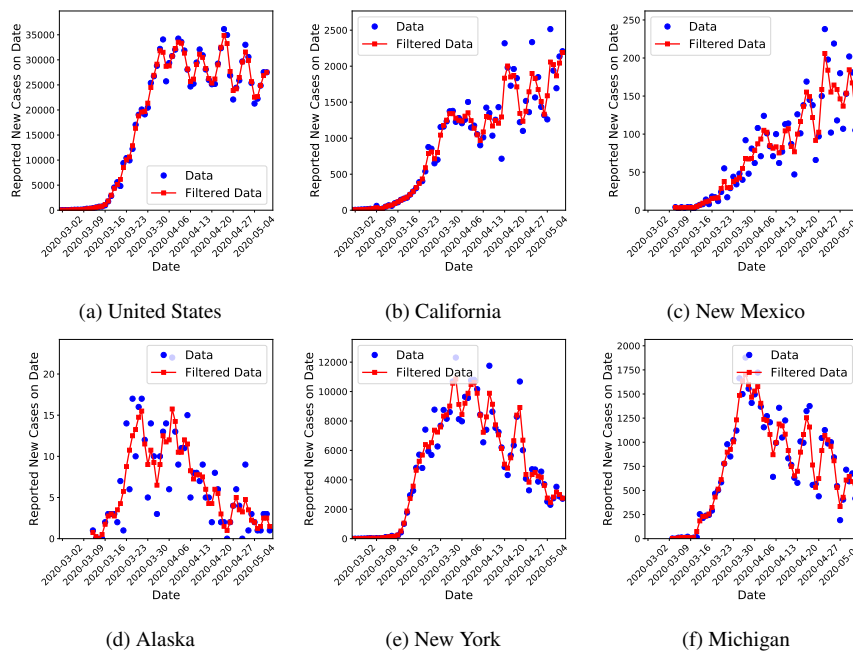


Fig. 18: Daily confirmed cases of COVID-19 aggregated the country and state level, shown in blue symbols, and the corresponding filtered data shown with red lines and symbols.

Fig. 18 shows data for all of the US (data extracted from [6]), and for 5 selected states (data extracted from [2]). The filtering approach, presented in §2.4, preserves the weekly scale variability observed for some of the datasets in this figure, and removes some of the large day to day variability observed for example in Alaska, in Fig. 18d.

Fig. 19 shows the data for several countries with a significant number of COVID-19 cases as of May 10, 2020. Similar to US and some of the US states, a weekly frequency trend can be observed superimposed on the overall epidemiological trend.

These trends are observed on the downward slope mostly, e.g. for Italy and Germany. When the epidemic is taking hold, it is possible that any higher frequency fluctuation is hidden inside the sharply increasing counts. Possible explanations include regional hot-spots flaring up periodically as well as expanded testing capabilities ramping-up over time.

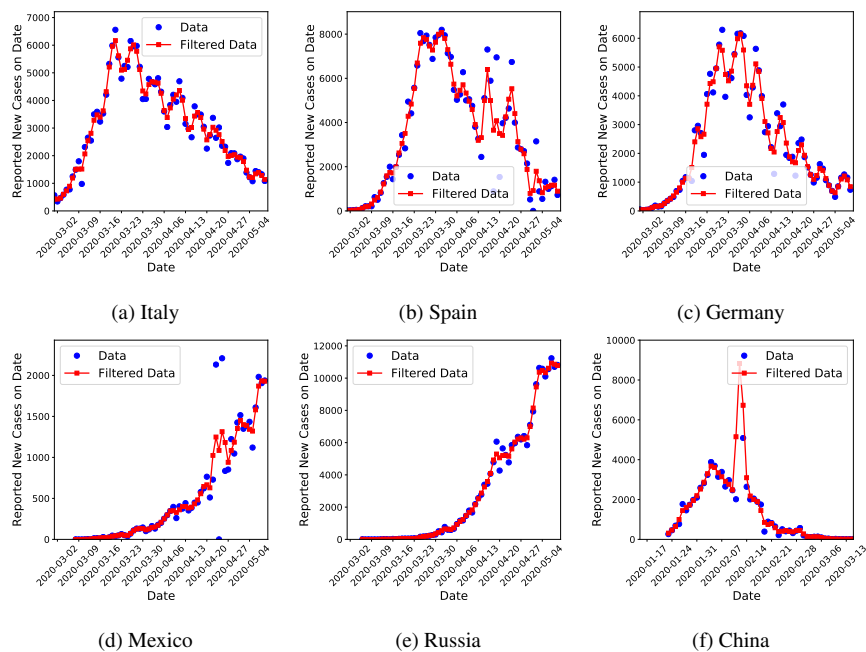


Fig. 19: Daily confirmed cases of COVID-19 for several countries shown in blue symbols and the corresponding filtered data shown with red lines and symbols. The large values for February 12 (14K) and 13 (5K) in China correspond to changes in how Chinese authorities defined “confirmed” cases. The data point for February 12 falls outside the y-axis in this figure.

We have also explored epidemiological models applied at regional scale. The left frame in Fig. 20 shows a set of counties in the Bay Area that were the first to issue the stay-at-home order on March 17, 2020. Two groups of counties in New Mexico are shown with red and blue in the right frame of Fig. 20. These regions displayed different disease dynamics, e.g. a shelter-in-place was first issued in the Bay Area on March 16, then extended to the entire state on March 19, while the new daily counts were much larger in the NW New Mexico compared to the central region. The daily counts, shown in Fig. 21 for these three regions was aggregated based on county data provided by [2].

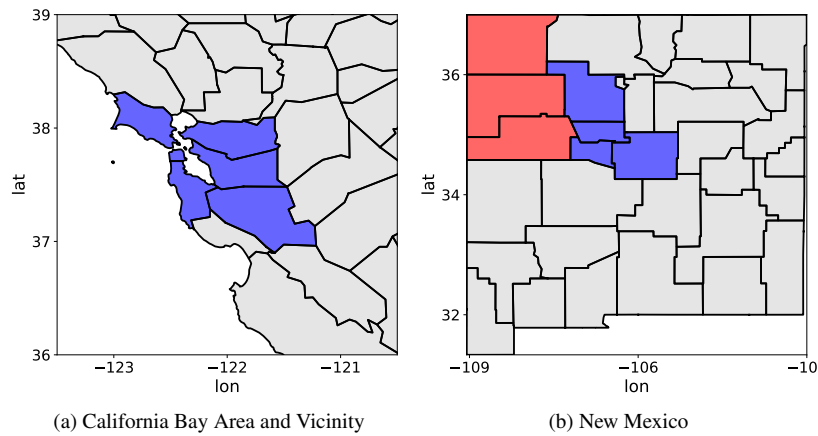


Fig. 20: Illustration of the set of counties aggregated together for the purpose of regional forecasts. Left frame: Marin, San Francisco, San Mateo, Contra Costa, Alameda, Santa Clara counties in the Bay Area shown in blue. Right frame: San Juan, McKinley, and Cibola counties in the north-west New Mexico shown in red, and Torrance, Valencia, Bernalillo, and Sandoval counties in central New Mexico shown in blue.

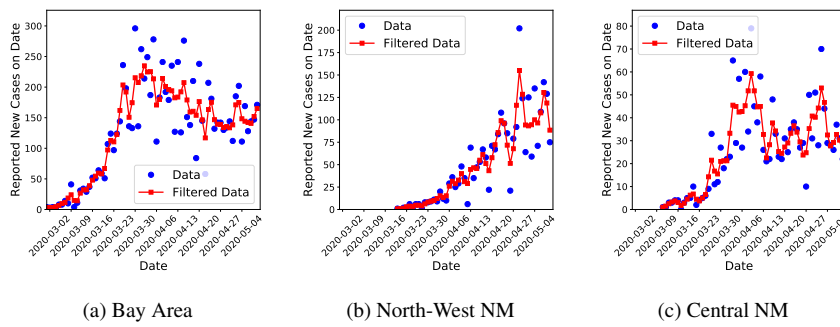


Fig. 21: Daily confirmed cases of COVID-19 shown in blue symbols and the corresponding filtered data shown with red lines and symbols for the three regions outlined in Fig. 20.