SANDIA REPORT
SAND2020-6563
Printed June, 2020

Characterization of Partially Observed Epidemics - Application to COVID-19

Cosmin Safta, Jaideep Ray, Erin Acquesta, Thomas Catanach, Kenny Chowdhary, Bert Debusschere, Edgar Galvan, Gianluca Geraci, Mohammad Khalil, and Teresa Portone

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods

ABSTRACT

This report documents a statistical method for the "real-time" characterization of partially observed epidemics. Observations consist of daily counts of symptomatic patients, diagnosed with the disease. Characterization, in this context, refers to estimation of epidemiological parameters that can be used to provide short-term forecasts of the ongoing epidemic, as well as to provide gross information for the time-dependent infection rate. The characterization problem is formulated as a Bayesian inverse problem, and is predicated on a model for the distribution of the incubation period. The model parameters are estimated as distributions using a Markov Chain Monte Carlo (MCMC) method, thus quantifying the uncertainty in the estimates.

The method is applied to the COVID-19 pandemic of 2020, using data at the country, provincial (e.g., states) and regional (e.g. county) levels. The epidemiological model includes a stochastic component due to uncertainties in the incubation period. This model-form uncertainty is accommodated by a pseudo-marginal Metropolis-Hastings MCMC sampler, which produces posterior distributions that reflect this uncertainty. We approximate the discrepancy between the data and the epidemiological model using Gaussian and negative binomial error models; the latter was motivated by the over-dispersed count data. For small daily counts we find the performance of the calibrated models to be similar for the two error models. For large daily counts the negative-binomial approximation is numerically unstable unlike the Gaussian error model.

Application of the model at the country level (for the United States, Germany, Italy, etc.) generally provided accurate forecasts, as the data consisted of large counts which suppressed the day-to-day variations in the observations. Further, the bulk of the data is sourced over the duration before the relaxation of the curbs on population mixing, and is not confounded by any discernible country-wide second wave of infections. At the state-level, where reporting was poor or which evinced few infections (e.g., New Mexico), the variance in the data posed some, though not insurmountable, difficulties, and forecasts were able to capture the data with large uncertainty bounds. The method was found to be sufficiently sensitive to discern the flattening of the infection and epidemic curve due to shelter-in-place orders after around 90% quantile for the incubation distribution (about 10 days for COVID-19). The proposed model was also used at a regional level to compare the forecasts for the central and north-west regions of New Mexico. Modeling the data for these regions illustrated different disease spread dynamics captured by the model. While in the central region the daily counts peaked in the late April, in the north-west region the ramp-up continued for approximately three more weeks.

ACKNOWLEDGMENT

The authors acknowledge the helpful feedback that Khachik Sargsyan and John Jakeman have provided on various aspects related to model inference and speed up of model evaluations. This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Application of this research to national analysis was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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-NA0003525.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. MODELING APPROACH

This work presents 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 during the same day. From this perspective, the forecasts presented in this report represent a lower bound on the actual number of people that are infected with COVID-19. A fraction of the population infected might exhibit minor or no symptoms at all and might not seek medical advice. Another set of cases might not seek medical advice even when showing symptoms. 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 §1.1 and an incubation rate model, described in §1.2. These models are combined, through a convolution presented in §1.3, into a forecast of number of cases that turn symptomatic daily. This collection of models are implemented in **COMBO** (**Co**vid-19 **M**odeling and **B**ayesian F**o**recast), which consists of a series of python scripts used to generate the results presented in this report.

1.1. Infection Rate Model

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.

The infection rate model is given by

$$f_\Gamma(t;k,\theta) = \theta^{-k} t^{k-1} \exp(-t/\theta) \big/ \Gamma(k) \tag{1.1.1}$$

where $f_\Gamma(t;k,\theta)$ it the probability density function (pdf) of the gamma distribution, with k and θ strictly positive. Fig. 1-1 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.

1.2. Incubation Model

We model the incubation distribution for COVID-19 using a lognormal distribution [12]. The nominal and 95% Confidence Interval values for the mean, μ, and standard deviation σ for the

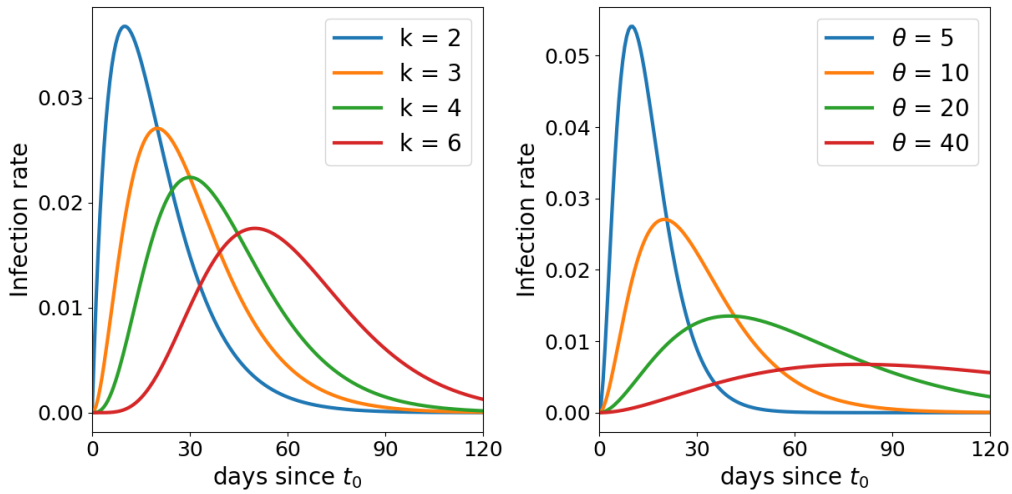


Figure 1-1. Infection rate models with fixed scale parameters $\theta = 10$ (left frame) and fixed shape parameter $k = 3$ (right frame).

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

Table 1-1. Nominal and 95% Confidence Interval (CI) values for the parameters of the incubation model (in Eq. (1.2.1)).

logarithm of the incubation model distribution are provided in Table 1-1. The pdf, f_{LN}, and cumulative distribution function (cdf), F_{LN}, of the lognormal distribution are given by

$$f_{LN} = \frac{1}{t\sigma\sqrt{2}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right), \quad F_{LN}(t;\mu,\sigma) = \frac{1}{2}\text{erfc}\left(-\frac{\log t - \mu}{\sigma\sqrt{2}}\right) \tag{1.2.1}$$

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 1-1. Let X be the logarithm of the incubation rate random variable. If X is approximately normally distributed, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a Student's t-distribution with n-degrees of freedom. To model the uncertainty in σ we assume that

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

has a χ^2 distribution with $(n-1)$ degrees of freedom. Here, S is a random variable corresponding to the sample standard deviation. While the data in [12] 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 distibutions assumed above to be narrower than the ranges provided in Table 1-1. 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. 1-2 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 then 85% of case completed the incubation process.

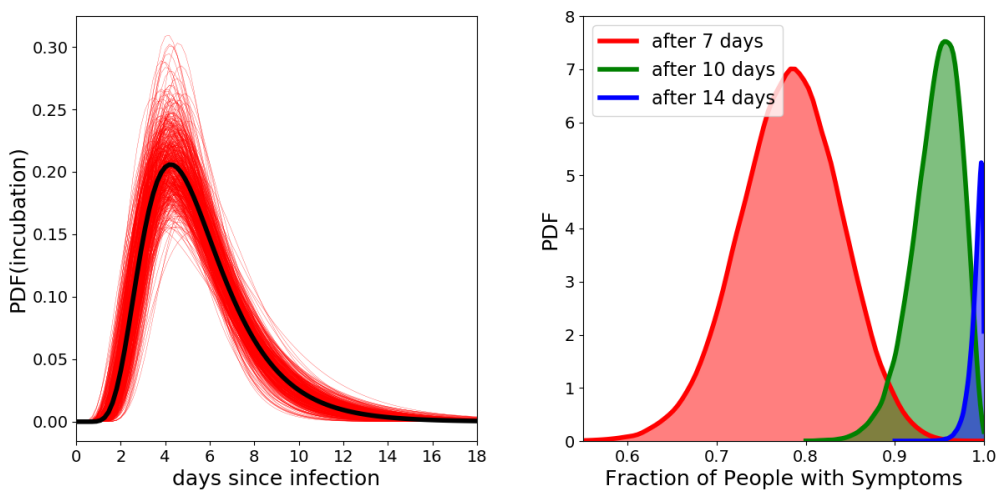


Figure 1-2. Probability densities for the incubation model (left frame) and fraction of people for which incubation ran its course after 7, 10, and 14 days respectively (right frame).

1.3. Daily Symptomatic Cases

With these assumptions the number of people infected *and* with completed incubation cycle at time t_i can be written as a convolution between the infection rate and the cumulative distribution function for the incubation distribution [7, 15, 17]

$$N_i = N \int_{t_0}^{t_i} f_\Gamma(\tau - t_0; k, \theta) F_{LN}(t_i - \tau; \mu, \sigma) d\,\tau \tag{1.3.1}$$

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

$$n_i = N_i - N_{i-1} = N \int_{t_0}^{t_i} f_\Gamma(\tau - t_0; k, \theta) \left(F^*_{LN}(t_i - \tau; \mu, \sigma) - F^*_{LN}(t_{i-1} - \tau; \mu, \sigma)\right) d\,\tau \tag{1.3.2}$$

where

$$F^*_{LN}(t;\mu,\sigma) = \begin{cases} 0 & \text{if } t < 0 \\ F_{LN}(t;\mu,\sigma) & \text{if } t \geq 0 \end{cases} \tag{1.3.3}$$

In Eq. (1.3.2), 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

$$F^*_{LN}(t_i - \tau;\mu,\sigma) - F^*_{LN}(t_{i-1} - \tau;\mu,\sigma) \approx (t_i - t_{i-1}) f_{LN}(t_i - \tau;\mu,\sigma) \tag{1.3.4}$$

leading to

$$n_i \approx N(t_i - t_{i-1}) \int_{t_0}^{(t_i+t_{i-1})/2} f_\Gamma(\tau - t_0;k,\theta) f_{LN}((t_i + t_{i-1})/2 - \tau;\mu,\sigma) d\,\tau \tag{1.3.5}$$

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. DATA

We have used several sources [1, 2, 3, 4, 5] 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. We found that for some states or regions the reported daily counts exhibited a significant amount of noise. This is caused 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 here a set of finite difference filters [10, 16] that preserve weekly or longer trends, and reduces noise, e.g. large day to day variability such as all cases for successive days being reported at the end of the time range.

$$\hat{\boldsymbol{y}} = F\boldsymbol{y},\ F = \mathbb{I} + (-1)^{n+1} 2^{-2n} \mathrm{D} \tag{2.0.1}$$

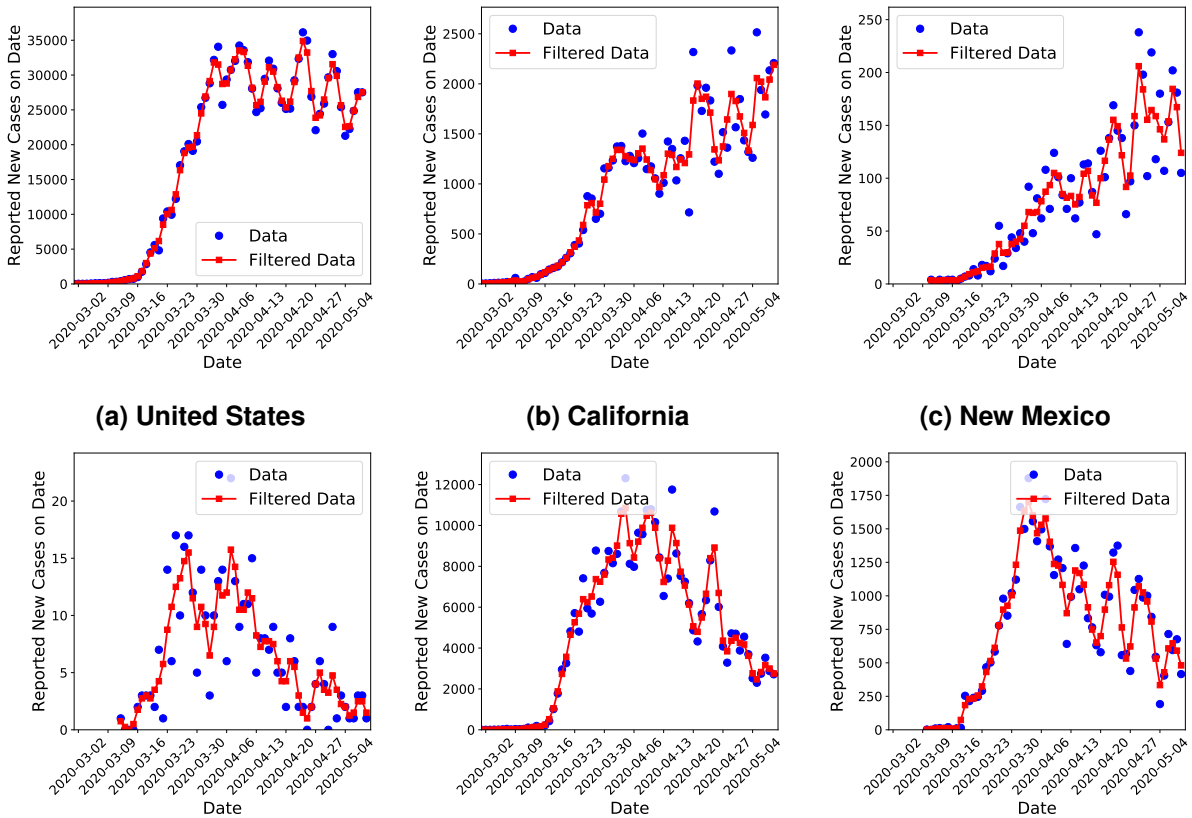
Here $\boldsymbol{y}$ is the original data, $\hat{\boldsymbol{y}}$ is the filtered data. Matrix $\mathbb{I}$ is the identity matrix and D is a band-diagonal matrix, e.g. triadiagonal for a n=2, i.e. a 2-nd order filter and pentadiagonal for a n=4, i.e. a 4-th order filter. We have compared 2-nd and 4-th order filters, and did not observe any significant difference between the filtered results. Reference [16] provides D matrices for filters up to 12-th order.

Fig. 2-1 shows data for all of the US (data extracted from [5]), and for 5 selected states (data extracted from [2]). The filtering approach preserves low wavenumber information, e.g. nearly 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. 2-1d.

Fig. 2-2 shows the data for several countries with a significant number of COVID-19 cases. Similar to US and some of the US states, a nearly weekly 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 periodic or higher frequency fluctuations are hidden inside the sharply increasing counts. We have not explored causes for these variations from the overall epidemic trends.

We have also explored epidemiological models applied at regional scale. The left frame in Fig. 2-3 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. 2-3. The daily counts, shown in Fig. 2-4 for these three regions was aggregated based on county data provided by [2].

For the remainder of this report 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 $\boldsymbol{y}$.



(a) United States **(b) California** **(c) New Mexico**

(d) Alaska **(e) New York** **(f) Michigan**

Figure 2-1. 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.

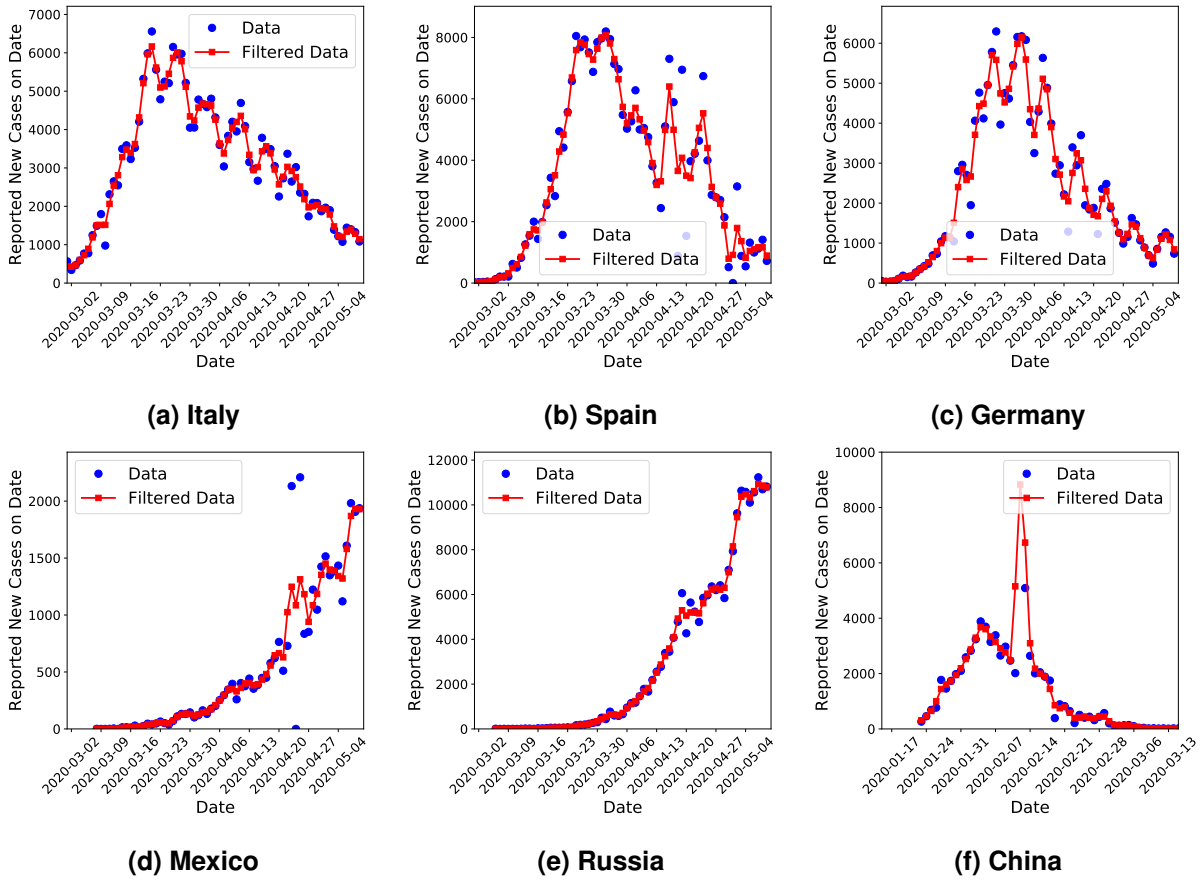
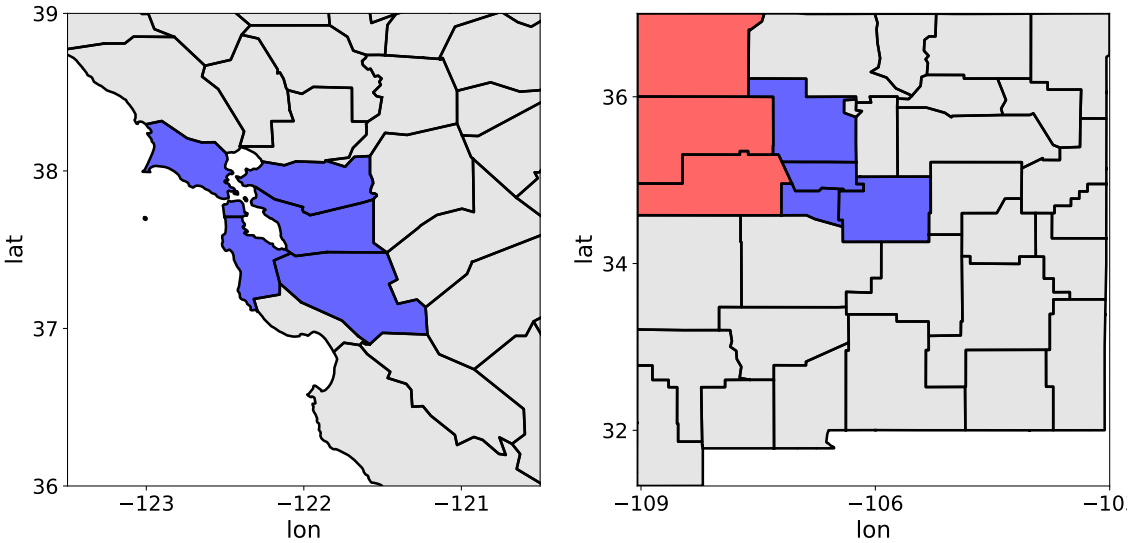
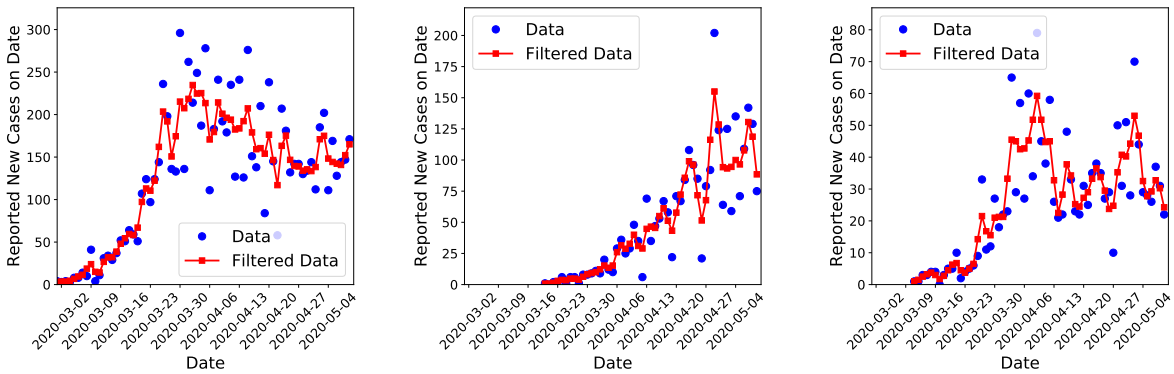


Figure 2-2. Daily confirmed cases of COVID-19 for several countries shown in blue symbols and the corresponding filtered data shown with red lines and symbols. The large values for February 12 (14K) and 13 (5K) in China correspond to changes in how Chinese authorities defined "confirmed" cases. The datapoint for February 12 falls outside the y-axis in this figure.



(a) California Bay Area and Vicinity **(b) New Mexico**

Figure 2-3. 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.



(a) Bay Area **(b) North-West NM** **(c) Central NM**

Figure 2-4. 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. 2-3.

3. MODEL INVERSION

Given data, $y(t)$, in the form of time-series of daily counts, as shown in Chapter 2, and the model predictions n_i for the number of new symptomatic counts daily, presented in Chapter 1, we write the discrepancy between the data and the model as

$$\boldsymbol{y} = \boldsymbol{n}(\Theta) + \varepsilon \tag{3.0.1}$$

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

$$\boldsymbol{y} = \{y(t_1), y(t_2), \ldots, y(t_D)\}, \quad \boldsymbol{n} = \{n_1(\Theta), n_2(\Theta), \ldots, n_D(\Theta)\}$$

where D is the number of data points and model parameters grouped as $\Theta = \{t_0, N, k, \theta\}$. Here 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 observations 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

$$p(\Theta|\boldsymbol{y}) \propto p(\boldsymbol{y}|\Theta)p(\Theta) \tag{3.0.2}$$

where $p(\Theta|\boldsymbol{y})$ is the posterior distribution we are seeking after observing the data $\boldsymbol{y}$, $p(\boldsymbol{y}|\Theta)$ is the likelihood of observing the data $\boldsymbol{y}$ 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 heterogenous 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 treated as Student's t and χ^2 random variables, respectively, as discussed in §1.2. Let us denote the underlying independent random variables by $\boldsymbol{\xi} = \{\xi_\mu, \xi_\sigma, \}$. The model prediction $n_i(\boldsymbol{\xi})$ is now a random variable induced by $\boldsymbol{\xi}$ plugged in Eq. (1.3.2), and $\boldsymbol{n}(\boldsymbol{\xi})$ is a random vector.

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

$$p(\boldsymbol{y}|\Theta) = \pi_{\boldsymbol{n}(\Theta)}(\boldsymbol{y}) = (2\pi)^{-D/2}|C_{\boldsymbol{n}}|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{n})^T C_{\boldsymbol{n}}^{-1}(\boldsymbol{y}-\boldsymbol{n})\right) \tag{3.1.1}$$

The covariance matrix $C_{\boldsymbol{n}}$ 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

$$C_{\boldsymbol{n},ii} = \sigma_i^2 = (\sigma_a + \sigma_m n_i(\Theta))^2 \tag{3.1.2}$$

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 positivity for the diagonal entries of $C_{\boldsymbol{n}}$. Under these assumptions, the MVN likelihood in Eq. (3.1.1) is written as a product of independent Gaussian densities

$$p(\boldsymbol{y}|\Theta) = \prod_{i=1}^{D}\pi_{n_i(\Theta)}(y_i) = (2\pi)^{-D/2}\prod_{i=1}^{D}(\sigma_a + \sigma_m n_i(\Theta))^{-1}\exp\left(-\frac{(y_i-n_i)^2}{2(\sigma_a+\sigma_m n_i(\Theta))^2}\right) \tag{3.1.3}$$

In §4 we will compare results obtained using only the additive part σ_a of Eq. 3.1.2 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 variance exceeds the mean [13]. 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

$$p(\boldsymbol{y}|\Theta) = \prod_{i=1}^{D}\pi_{n_i(\Theta)}(y_i) = \prod_{i=1}^{D}\binom{y_i+\alpha-1}{\alpha-1}\left(\frac{\alpha}{\alpha+n_i}\right)^{\alpha}\left(\frac{n_i}{\alpha+n_i}\right)^{y_i} \tag{3.1.4}$$

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

$$\binom{y_i+\alpha-1}{\alpha-1} = \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)\Gamma(y_i+1)} \tag{3.1.5}$$

is the binomial coefficient.

For the *stochastic incubation model* the likelihood components are not analytical anymore since the model involves a non-linear convolution.

$$p(\boldsymbol{y}|\Theta) = \pi_{\boldsymbol{n}(\Theta),\boldsymbol{\xi}}(\boldsymbol{y}) \tag{3.1.6}$$

The likelihood evaluations are now a density evaluation process; for every sample of Θ, we sample $\boldsymbol{\xi}$ to construct a set of model samples $\boldsymbol{n}$. These samples are then augmented with random samples from the assumed discrepancy ε and the resulting samples are used to estimate $\pi_{n_i(\Theta),\boldsymbol{\xi}}(y_i)$ via a density estimation procedure.

$$\pi_{n_i(\Theta),\boldsymbol{\xi}}(y_i) = \frac{1}{S}\sum_{k=1}^{S} \mathscr{K}\left(n_i^{(k)}(\Theta) - y_i\right) \tag{3.1.7}$$

where S is the number of samples, $n_i^{(k)}(\Theta)$ is the model prediction for day i corresponding to sample $\boldsymbol{\xi}^{(k)}$, and $\mathscr{K}$ is the kernel centered at this sample. Most density estimation algorithms employ Gaussian kernels. This approach has to be repeated for all posterior samples and is prohibitive for most practical applications. The high-dimensionality of the data, in this case D increases from around 14 (days of data) to about 60, presents an additional challenge to the density estimation procedure. To alleviate this challenge we approximate the likelihood as a product of marginal discrepancies for each data sample

$$\pi_{\boldsymbol{n}(\Theta),\boldsymbol{\xi}}(\boldsymbol{y}) = \prod_{i=1}^{D} \pi_{n_i(\Theta),\boldsymbol{\xi}}(y_i) = \prod_{i=1}^{D}\left(\frac{1}{S}\sum_{k=1}^{S} \mathscr{K}\left(n_i^{(k)}(\Theta) - y_i)\right)\right) \tag{3.1.8}$$

This approach will be employed for both discrepancy models, Gaussian and negative-binomial, Eqs. (3.1.3) and (3.1.4).

3.2. Posterior Sampling

A Markov Chain Monte Carlo (MCMC) algorithm is used to sample from the posterior density $p(\Theta|\boldsymbol{y})$. 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 [9]. We have also explored additional algorithms, including transitional MCMC (TMCMC) [11] as well as ensemble samplers [8] 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.

Posterior Sampling for the Stochastic Incubation Model We will employ a pseudo-marginal MCMC algorithm [6] to circumvent some of the computational cost related to estimating the marginal densities in Eq. (3.1.8) at every MCMC step. The Metropolis-Hastings algorithm draws new Θ samples from a proposal $q(\cdot|\Theta_i)$ then accepts the sample with probability

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

where $p(\Theta_i|\boldsymbol{y})$ and $p(\Theta_{i+1}|\boldsymbol{y})$ 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, $q(\Theta_i|\Theta_{i+1}) = q(\Theta_{i+1}|\Theta_i)$.

When the likelihood is computationally expensive the pseudo-marginal MCMC replaces the likelihood evaluations in Eq. (3.1.8) with an unbiased distribution $\hat{\pi}_{\boldsymbol{n}(\Theta)}(\boldsymbol{y})$, with $\mathbb{E}[\hat{\pi}_{\boldsymbol{n}(\Theta)}(\boldsymbol{y})] = \pi_{\boldsymbol{n}(\Theta)}(\boldsymbol{y})$ at every Θ. For the work presented here, at each MCMC step we draw a random sample $\boldsymbol{\xi}$ from its distribution, and then we estimate the likelihood in a way similar to the deterministic incubation model, in Eqs. (3.1.3) or (3.1.4).

3.3. Predictive Assessment

We will employ both pushed-forward distributions and Bayesian posterior-predictive distributions [14] to assess the predictive skill of the proposed statistical framework to model the COVID-19 disease spread. The schematic in Eq. (3.3.1) illustrates the process to generate push-forward posterior estimates

$$p(\Theta|\boldsymbol{y}) \rightarrow \{\underbrace{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(m)}}_{\text{Posterior Samples}}\} \xrightarrow{\boldsymbol{n}(\Theta)} \{\boldsymbol{y}^{\text{rep},(1)}, \boldsymbol{y}^{\text{rep},(2)}, \ldots, \boldsymbol{y}^{\text{rep},(m)}\} \rightarrow p_{\text{pf}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y}). \tag{3.3.1}$$

Here, $\boldsymbol{y}^{\text{rep}}$ denotes hypothetical or "replicated" data $\boldsymbol{y}$ and $p_{\text{pf}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y})$ denotes the push-forward probability density of the hypothetical data $\boldsymbol{y}^{\text{rep}}$ conditioned on the observed data $\boldsymbol{y}$. We start with samples from the posterior distribution $p(\Theta|\boldsymbol{y})$. These samples are readily available from the MCMC exploration of the parameter space. Typically we subsample the MCMC chain to about 5-10K samples that will be used to generate push-forward statistics. Using these samples, we evaluate the epidemiological model and collect the resulting $\boldsymbol{y}^{\text{rep}} = \boldsymbol{n}(\Theta)$ samples that correspond to the push-forward posterior distribution $p_{\text{pf}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y})$.

The pushed-forward posterior does not account for the discrepancy between the data $\boldsymbol{y}$ and the model predictions $\boldsymbol{n}$, subsumed into the definition of the error model ε presented in Eqs. (3.1.3) and (3.1.4). The Bayesian posterior-predictive distribution, defined in Eq. (3.3.2) is computed by marginalization of the likelihood over the posterior distribution of model parameters Θ:

$$p_{\text{pp}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}^{\text{rep}}|\Theta) p(\Theta|\boldsymbol{y}) d\Theta. \tag{3.3.2}$$

In practice, we estimate $p_{\text{pp}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y})$ through sampling, because analytical estimates are not usually available. The sampling workflow is similar to the one shown in Eq. (3.3.1). After the model evaluations $\boldsymbol{y} = \boldsymbol{n}(\Theta)$ 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 $p_{\text{pp}}(\boldsymbol{y}^{\text{rep}}|\boldsymbol{y})$.

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 $\boldsymbol{y}^{\text{rep}}$ that will be compared against $\boldsymbol{y}$, the data used to infer the model parameters. The same functional form can be used to generate statistics for $\boldsymbol{y}^{\text{rep}}$ 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.

Fig. 4-1 shows a schematic of the workflow that assembles the modeling components described in §1-§3. First, we will compare the predictive capabilities of these models for several modeling

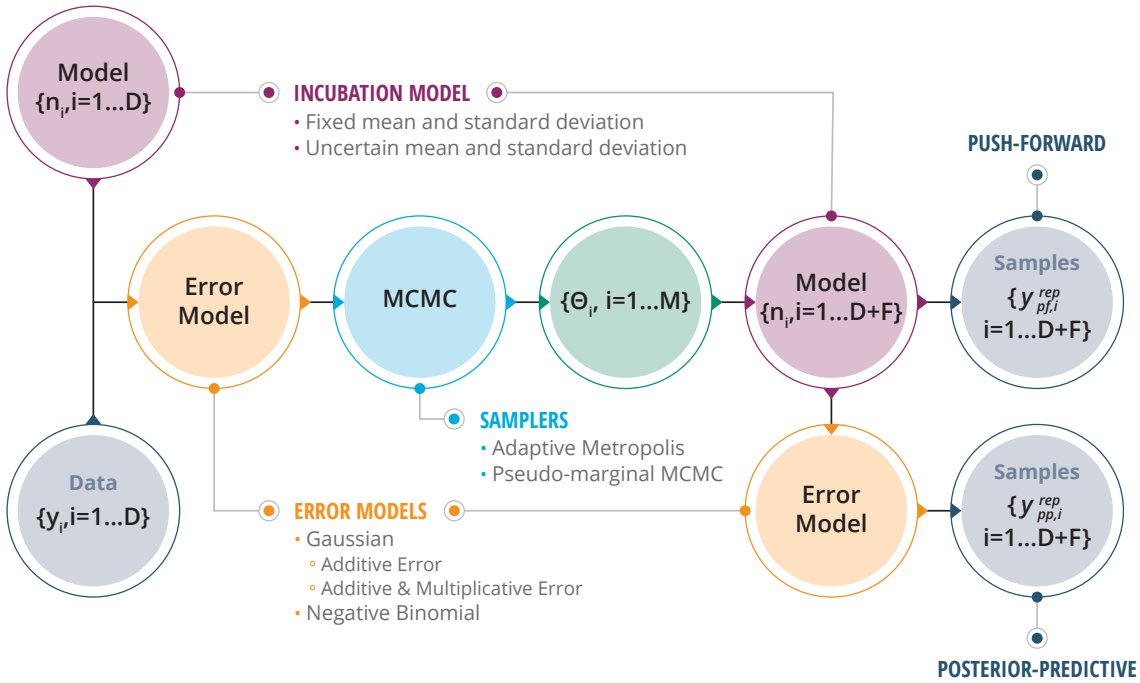


Figure 4-1. Epidemiological model inference and forecast workflow. The "Model" circle encompasses the convolution between the infection rate and the incubation model described in §1 and the "Error Model" circle illustrates the choices for the discrepancy between the model and the data presented in §3.1.

choices:

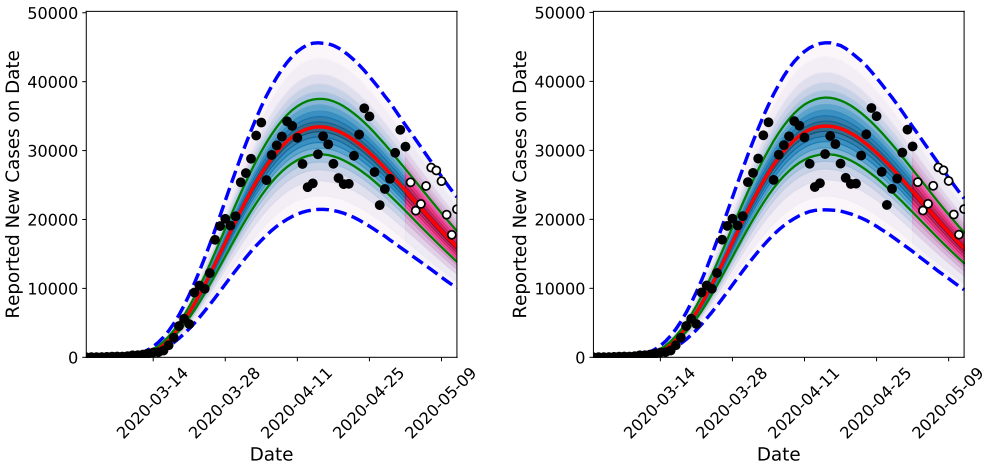
- §4.2: Deterministic vs. Stochastic Incubation Models
- §4.3: Additive vs Additive+Multiplicative Error Models
- §4.4: Gaussian vs Negative Binomial Likelihood Models

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 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 quantiles 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. Deterministic vs Stochastic Incubation Models



(a) Deterministic Incubation Rate **(b) Stochastic Incubation Rate**

Figure 4-2. 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.

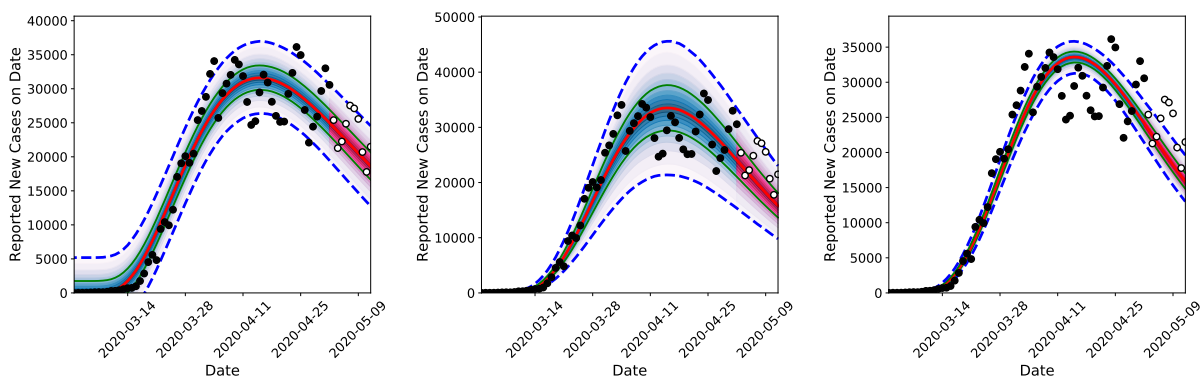
We start the analysis with an assessment of the impact of the choice of incubation model on the model prediction. First we ran our model using the log-normal incubation model with mean and standard deviation fixed at their nominal values in Table 1-1. We then used the same dataset to calibrate the epidemiological model which employs an incubation rate with uncertain mean and standard deviation as described in §1.2. These results are labeled "Deterministic" and "Stochastic", respectively, in Fig. 4-2. This figure shows results based on data corresponding to

the United States. The choice of deterministic vs stochastic incubation models produce very similar output.

The results shown in the right frame of Fig. 1-2 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 and, for the results presented below we employ an incubation model with nominal parameters.

4.3. Additive vs Additive-Multiplicative Error Models

Next, we explore results based on either additive error (AE) or additive+multiplicative error (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. (3.1.2). For AE we only infer σ_a while for A+ME we infer both σ_a and σ_m. The AE results in Fig. 4-3a are based on the same dataset as the A+ME results in Fig. 4-3b. Both



(a) AE Posterior-Predictive **(b) A+ME Posterior Predictive** **(c) A+EM Push Forward**

Figure 4-3. Posterior predictive forecast for (a) additive error (AE) models and (b) additive+multiplicative error (A+EM) models, and (c) push forward posterior forecasts for A+EM, using data aggregated for all of the United States. The middle frame is the same as the right frame in Fig. 4-2 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. 4-3a, 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. 4-3c, and the posterior-predictive data, in Fig. 4-3b. 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

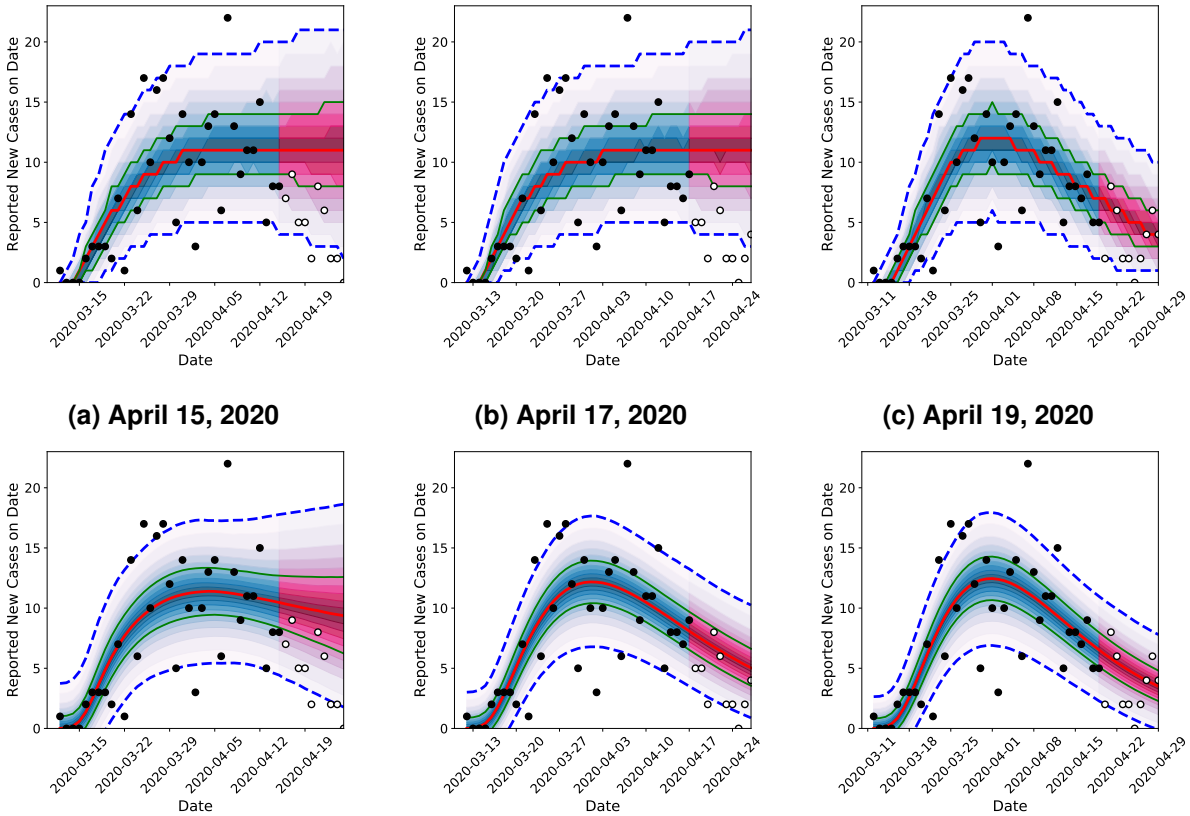
The negative-binomial distribution is used commonly in epidemiology to model overly dispersed data, e.g. in cases where the variance exceeds the mean [13]. We also observe similar trends in some of the COVID-19 datasets. Fig. 4-4 shows results based on data for Alaska. The results based on the two error 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.

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. 4-5 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 dispersion. To capture this increased noise, we switched from an AE model to A+ME model, with results shown in Fig. 4-6.



(a) April 15, 2020 **(b) April 17, 2020** **(c) April 19, 2020**

(d) April 15, 2020 **(e) April 17, 2020** **(f) April 19, 2020**

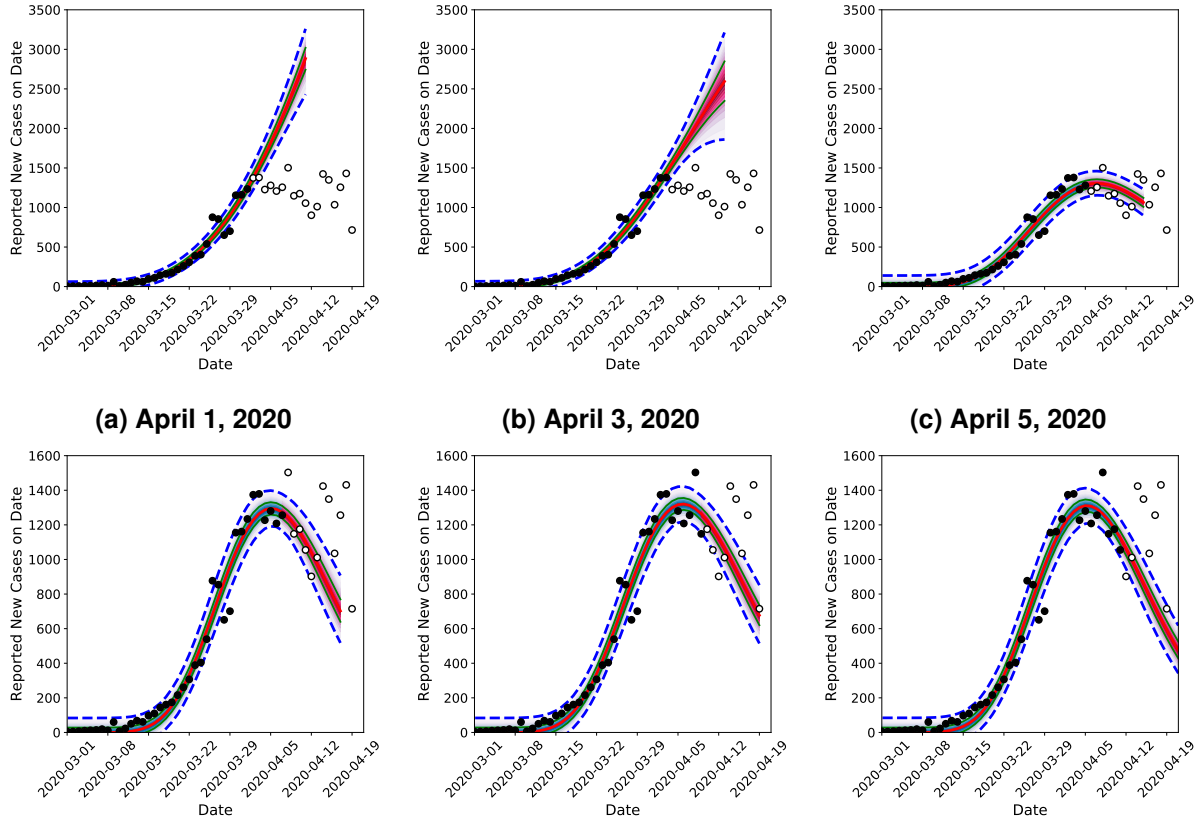
Figure 4-4. Posterior-predictive forecasts for Alaska, using negative binomial likelihood (top row) and additive/multiplicative Gaussian likelihood (bottom row). Symbols and colors annotations are described in §4.1.

4.5.2. Example of Dynamics at Regional Scale: New Mexico

Figs. 4-7 and 4-8 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. 2-3b. 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.

4.5.3. Moving target for US

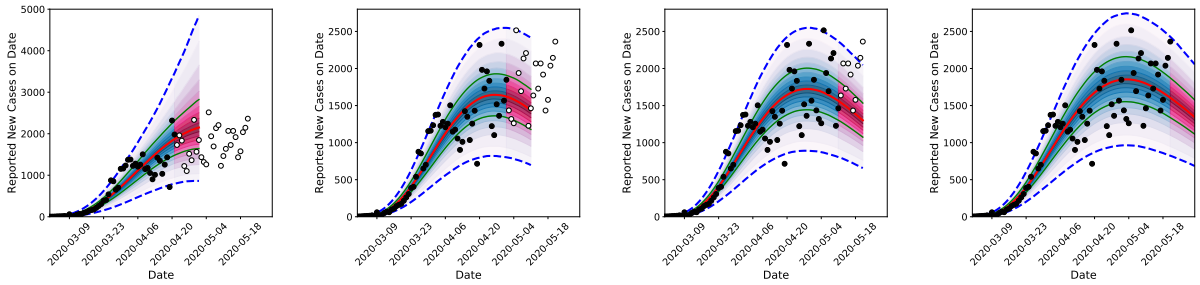
This section discusses an analysis of the aggregate data from all US states. The posterior-predictive results shown in Fig. 4-9a—4-9d suggest the peak in the number of daily cases was reached around mid-April. Nevertheless the model had to adjust the downward slope as



(a) April 1, 2020 **(b) April 3, 2020** **(c) April 5, 2020**

(d) April 7, 2020 **(e) April 9, 2020** **(f) April 11, 2020**

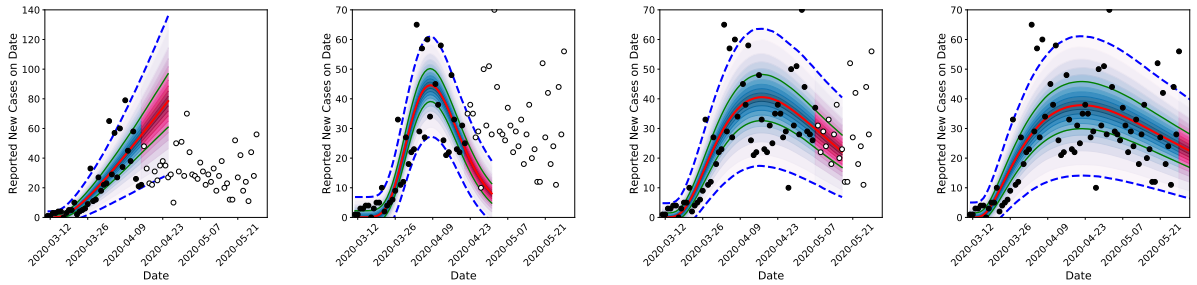
Figure 4-5. Posterior-predictive forecasts for California, based on additive error models using data available on (a) April 1, 2020 through (f) April 11. Symbols and colors annotations are described in §4.1.



(a) April 21, 2020 **(b) May 1, 2020** **(c) May 11, 2020** **(d) May 21, 2020**

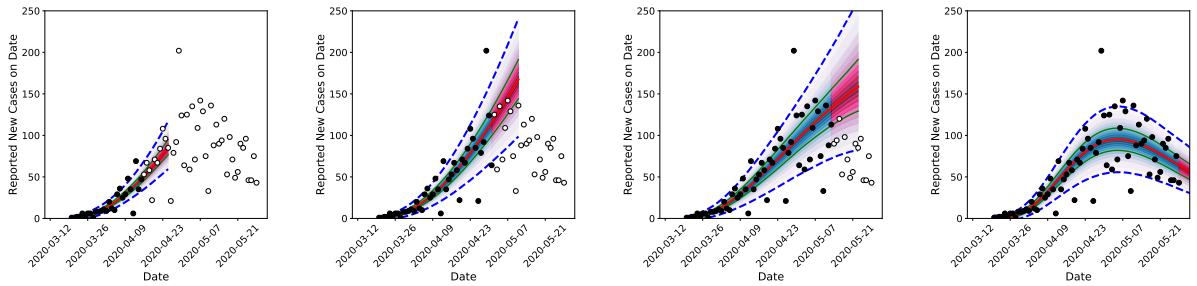
Figure 4-6. Posterior-predictive forecasts for California, based on additive/multiplicative error models using data available on (a) April 21, 2020 through (d) May 21, 2020. Symbols and colors annotations are described in §4.1.

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



(a) April 15, 2020 **(b) April 21, 2020** **(c) May 7, 2020** **(d) May 28, 2020**

Figure 4-7. Posterior-predictive forecasts for New Mexico central region, corresponding to counties highlighted with blue in Fig. 2-3b. Symbols and colors annotations are described in §4.1.



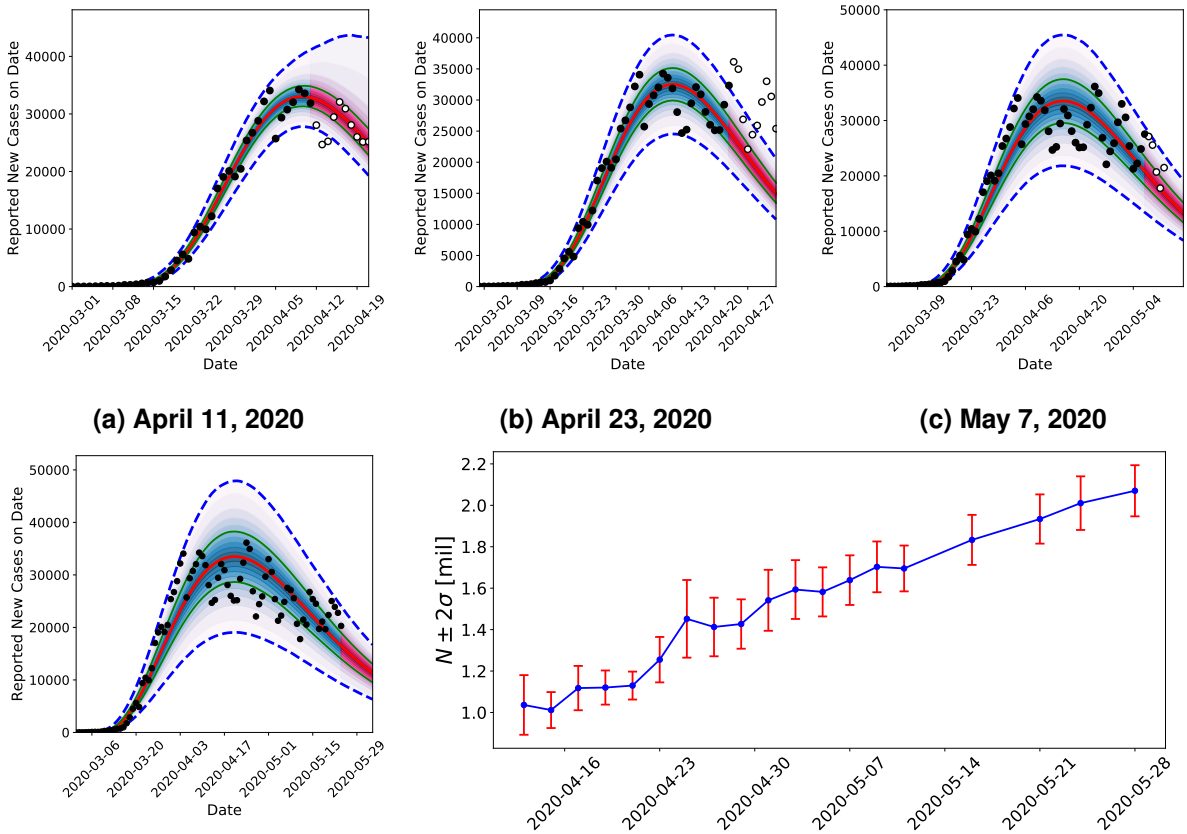
(a) April 15, 2020 **(b) May 1, 2020** **(c) May 13, 2020** **(d) May 28, 2020**

Figure 4-8. Posterior-predictive forecasts for New Mexico north-west region, corresponding to counties highlighted with red in Fig. 2-3b. Symbols and colors annotations are described in §4.1.

show in Fig. 4-9e.

4.5.4. Sequence of Forecasts for Other Countries

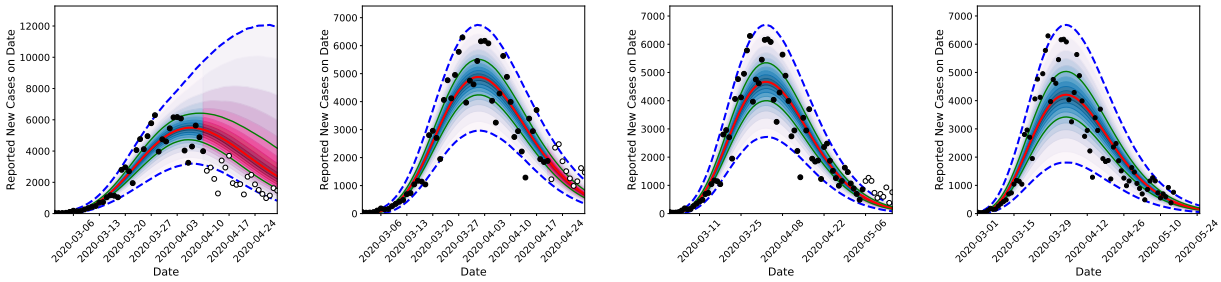
We conclude our analysis of the proposed epidemiological model with available daily symptomatic cases pertaining to Germany, Italy, and Spain, in Figs. 4-10, 4-12, and 4-11, respectively. For Germany, the uncertainty range increases while the epidemic is winding down, as the data has a relatively large 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. 2-2b. 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. 4-12, exhibit an upward shift observed around April 10-20, similar to data for Spain above. The subsequent forecasts display narrower uncertainty bands compared to other similar forecasts above, possibly due to the



(a) April 11, 2020 (b) April 23, 2020 (c) May 7, 2020

(d) May 24, 2020 (e) Total number of cases

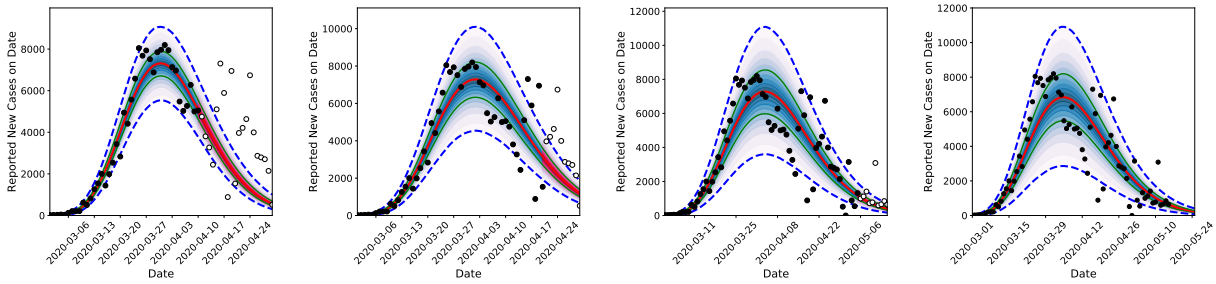
Figure 4-9. (a-d) Posterior-predictive forecasts for US, based on additive/multiplicative error models and (f) Total number of cases N. Symbols and colors annotations for (a)-(d) are described in §4.1.



(a) April 10, 2020 (b) April 20, 2020 (c) May 5, 2020 (d) May 15, 2020

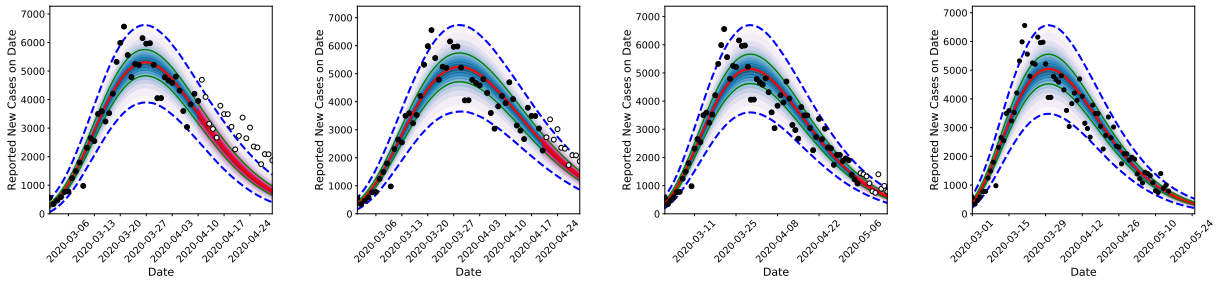
Figure 4-10. Posterior-predictive forecasts for Germany, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1.

absence of hotspots and/or regular data reporting.



(a) April 10, 2020 **(b) April 20, 2020** **(c) May 5, 2020** **(d) May 15, 2020**

Figure 4-11. Posterior-predictive forecasts for Spain, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1.



(a) April 10, 2020 **(b) April 20, 2020** **(c) May 5, 2020** **(d) May 15, 2020**

Figure 4-12. Posterior-predictive forecasts for Italy, based on additive/multiplicative error models. Symbols and colors annotations are described in §4.1.

5. SUMMARY

This report illustrates the performance of a method for producing short-term forecasts (with a forecasting horizon of about 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 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 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 as well. 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 models inference 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.

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 Coronavirus Pandemic. https://www.worldometers.info/coronavirus. Accessed: 2020-05-10.

[4] 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.

[5] 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.

[6] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.

[7] R. Brookmeyer and M. H. Gail. 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.

[8] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1):65–80, 2010.

[9] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[10] C. A. Kennedy and M. H. Carpenter. Several new numerical methods for compressible shear-layer simulations. *Applied Numerical Mathematics*, 14(4):397 – 433, 1994.

[11] M. Khalil, J. Lao, C. Safta, and H.N. Najm. Transitional Markov Chain Monte Carlo Sampler in UQTk. Technical Report SAND2020-3166, Sandia National Laboratories, February 2020.

[12] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, 2020.

[13] J. O. Lloyd-Smith. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. *PLOS ONE*, 2(2):1–8, 02 2007.

[14] S. M. Lynch and B. Western. Bayesian posterior predictive checks for complex models. *Sociological Methods and Research*, 32(3):301–335, 2004.

[15] J. Ray and S. Lefantzi. Deriving a model for influenza epidemics from historical data. Technical Report SAND2011-6633, Sandia National Laboratories, September 2011.

[16] Jaideep Ray, Christopher A. Kennedy, Sophia Lefantzi, and Habib N. Najm. Using High-Order Methods on Adaptively Refined Block-Structured Meshes: Derivatives, Interpolations, and Filters. *SIAM Journal on Scientific Computing*, 29(1):139–181, 2007.

[17] C Safta, J. Ray, K. Sargsyan, S. Lefantzi, K. Cheng, and D. Crary. Real-time Characterization of Partially Observed Epidemics using Surrogate Models. Technical Report SAND2011-6776, Sandia National Laboratories, September 2011.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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-NA0003525.