HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2024 September 07.

Published in final edited form as:

Ann Appl Stat. 2023 September ; 17(3): 1801–1819. doi:10.1214/22-aoas1671.

REAL-TIME MECHANISTIC BAYESIAN FORECASTS OF COVID-19 MORTALITY

Graham C. Gibson[1,a], **Nicholas G. Reich**[1,b], **Daniel Sheldon**[2,c]

[1]School of Public Health and Health Sciences, University of Massachusetts Amherst

[2]College of Information and Computer Sciences, University of Massachusetts Amherst

Abstract

The COVID-19 pandemic emerged in late December 2019. In the first six months of the global outbreak, the U.S. reported more cases and deaths than any other country in the world. Effective modeling of the course of the pandemic can help assist with public health resource planning, intervention efforts, and vaccine clinical trials. However, building applied forecasting models presents unique challenges during a pandemic. First, case data available to models in real time represent a nonstationary fraction of the true case incidence due to changes in available diagnostic tests and test-seeking behavior. Second, interventions varied across time and geography leading to large changes in transmissibility over the course of the pandemic. We propose a mechanistic Bayesian model that builds upon the classic compartmental susceptible–exposed–infected–recovered (SEIR) model to operationalize COVID-19 forecasting in real time. This framework includes nonparametric modeling of varying transmission rates, nonparametric modeling of case and death discrepancies due to testing and reporting issues, and a joint observation likelihood on new case counts and new deaths; it is implemented in a probabilistic programming language to automate the use of Bayesian reasoning for quantifying uncertainty in probabilistic forecasts. The model has been used to submit forecasts to the U.S. Centers for Disease Control through the COVID-19 Forecast Hub under the name MechBayes. We examine the performance relative to a baseline model as well as alternate models submitted to the forecast hub. Additionally, we include an ablation test of our extensions to the classic SEIR model. We demonstrate a significant gain in both point and probabilistic forecast scoring measures using MechBayes, when compared to a baseline model, and show that MechBayes ranks as one of the top two models out of nine which regularly submitted to the COVID-19 Forecast Hub for the duration of the pandemic, trailing only the COVID-19 Forecast Hub ensemble model of which which MechBayes is a part.

[a] gcgibson@lanl.gov. [b] nick@umass.edu. [c] sheldon@cs.umass.edu.

SUPPLEMENTARY MATERIAL

Code (DOI: 10.1214/22-AOAS1671SUPPA; .zip). This supplement provides the code used to build and run MechBayes as well as produce submission files as submitted to the CDC.

Appendix (DOI: 10.1214/22-AOAS1671SUPPB; .pdf). This supplement provides further details of the priors used on all parameters governing both the differential equation and observation models. Additionally, it provides further details of the initialization of the mechanistic model, including priors on initial compartmental values and computation of incident cases from MechBayes.

Keywords

COVID-19 mortality forecasting; compartmental models; Bayesian differential equations

1. Introduction.

The emergence of COVID-19 in early 2020 led to the largest pandemic in over a century. Understanding the future trajectory of the pandemic can help decision-makers prepare for and consequently diminish the impact in terms of healthcare and economic burden. Forecasts of incident and cumulative deaths due to COVID-19 help in resource allocation and reopening strategies (Ray et al. (2020)). Forecasts provide important data to decision-makers and the general public and can improve situational awareness of current trends and how they will likely evolve in coming weeks.

Infectious disease forecasting, at the time horizon of up to four weeks in the future, has benefited public health decision makers during annual influenza outbreaks (Lutz et al. (2019), Myers et al. (2000)). However, many forecasts of endemic, seasonal diseases, such as influenza, rely on ample historical data to look for patterns in the training data that can be projected forward into the future for extrapolation. In an emerging pandemic situation, models must be able to fit to limited data. Mechanistic models are a natural framework for modeling and forecasting in a limited data scenario, such as COVID-19. These directly model the transmission of the disease through the population and can be fit to public health surveillance data with relatively few parameters. In a Bayesian context, compartmental models can lever-age distributional estimates of parameters from smaller epidemiological studies to inform population level dynamics.

Our work is based on compartmental models, which are classical mechanistic models for disease transmission, that were first introduced by Kermack and McKendrick (1927). These assume that, at any given time, each individual is in one of a mutually exclusive set of compartments, typically either the susceptible, exposed, infected, or recovered compartment. A model is specified by setting the rates of flow of individuals between compartments. While these models have been used since their inception in the early 20th century, the COVID-19 pandemic represents a unique opportunity to explore their operational forecasting properties in real time at local, national, and global scales, for an emerging pathogen that, unlike influenza, does not have years of data on which models can train. This work details a fully operational model (which is available as a python package[1]) that performed second overall in the U.S. Centers for Disease Control COVID-19 forecasting initiative, trailing only the COVID Hub ensemble model which included MechBayes as a component model.

We develop a mechanistic Bayesian model that tailors compartmental models to the operational needs of making one- to four-week ahead forecasts of incident deaths for COVID-19. Since the primary goal is to parsimoniously forecast an observable quantity,

[1] https://github.com/dsheldon/mechbayes

estimating internal parameters of the model, many of which are poorly determined or not identifiable from the available data (Korolev (2021)), is not an explicit focus or prerequisite of our work. We distinguish this setup from longer-term scenario projection models, which require well identified epidemiological parameters that can be set to counterfactual values under different scenarios, such as an increase or decrease in intervention levels (Pei, Kandula and Shaman (2020), Borchering et al. (2021)). Scenario projection models are often based on similar foundations but require different adaptations than those needed for real-time forecasting.

Our model is tailored to the particular needs and data availability of COVID-19. The compartmental model jointly models infections and deaths and uses records of both incident cases and deaths—the two most widely available COVID-19 surveillance measurements—for model fitting. Our model includes components to model changes over time in both the dynamics and the detection of COVID-19. In particular, transmission rates have changed significantly due to the addition and removal of control measures, such as social distancing, lockdown, and mask use, while infection reporting rates have changed due to significant changes in the availability of diagnostic testing (Catching et al. (2021), Flaxman et al. (2020), Lau et al. (2021)).

There have been many Bayesian differential equation models explored in an infectious disease context, particularly with COVID-19 (Johndrow et al. (2020), Yang and Lee (2020), Zhuang et al. (2013), Mbuvha and Marwala (2020), Zhuang and Cressie (2014), Grinsztajn et al. (2021)). Many of the models submitted to the CDC COVID-19 forecasting initiative are based on underlying compartmental models (Karlen (2020), Ray et al. (2020)). However, the heterogeneity of model performance (Cramer et al. (2021a)) shows that not all compartmental models are created (or implemented) equally. While both the theory behind the models and the ability to fit them through Bayesian methods has been established, the particular choices modelers face when adapting the theory to real data has consequences for forecast accuracy. In particular, MechBayes uses a flexible transmission model, flexible link between cases and deaths, and employs efficient fully Bayesian inference. The ability to build a modular probabilistic model with complex components and automatically obtain efficient inference procedures is a testament to recent advances in algorithms and software packages largely driven by advances in Markov chain Monte Carlo methodology, autodifferentiation, and automatic variable transformations for latent variables into an unconstrained parameter space (Uber (2020)). What separates our model from previous work is its practical application and demonstrated success in a real-time prospective CDC COVID-19 mortality forecasting initiative (Cramer et al. (2021a)).

We demonstrate the success of our model by showing that forecasts submitted to the CDC (under the model name “MechBayes”) via the COVID-19 Forecast Hub outperform a baseline probabilistic forecast model described in Cramer et al. (2021a). We additionally show (and point to independent evaluations that support the conclusion) that MechBayes is one of the top performing models out of those submitted to the COVID-19 Forecast Hub. Finally, we quantify the features of MechBayes that improve the model over a traditional compartmental model via an ablation study.

2. Methods.

2.1. Data.

In this analysis we used confirmed case counts and deaths for the 50 U.S. states and the District of Columbia, as reported by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) (Dong, Du and Gardner (2020)). The data set reports the incident number of confirmed cases and deaths for each location at a daily frequency starting in early 2020. As noted in Krantz and Rao (2020), COVID-19 cases are underreported, with the fraction of all infections reported as cases for the U.S. estimated at 20–30% (Russel et al. (2020)). The fraction of all infections reported has also changed dramatically over the course of the epidemic (Rahmandad, Lim and Sterman (2020)).

2.2. Forecast targets.

We made probabilistic forecasts for one to four weeks ahead incident and cumulative deaths for all geographies. An individual forecast distribution is represented by a set of 23 quantiles, $\mathbb{Q} = \{0.01,\ 0.05,\ 0.10,\ \ldots,\ 0.90,\ 0.95,\ 0.99\}$, with the median (0.50 quantile) representing the point forecast. While forecasts are made at the daily scale, we aggregate them to the weekly scale by summing incident death forecasts from the first forecasted Sunday through the following Saturday. We evaluate only forecasts of incident deaths which is the primary modeled quantity; forecasts for cumulative deaths are created by accumulating forecasted incident deaths.

2.3. Mechanistic Bayesian model.

Compartmental models have been used to effectively model and forecast disease in nonpandemic situations both retrospectively and in real time. These include complex compartmental models for real-time influenza forecasting (Shaman and Karspeck (2012), Osthus et al. (2017), Ong et al. (2010)) and a retrospective model evaluation of the 1918 influenza pandemic (Hall et al. (2007)). Compartmental models have been used for both inference and forecasting not just in respiratory disease but in Ebola (Lekone and Finkenstädt (2006)), measles (Bokler (1993)), dengue (Syafruddin and Noorani (2012)), and a wide variety of other communicable diseases.

Compartmental models have also been adopted into a Bayesian framework before, including both stochastic disease dynamics and deterministic dynamics (Hotta (2010), Dukic, Lopes and Polson (2012)). Nonparametric transmissibility was included in a Bayesian SEIR model to study Ebola by Frasso and Lambert (2016). Time-varying transmissibility has also been studied in the frequentist setting using complex nonparametric functions (Smirnova, deCamp and Chowell (2019)). Many efforts have been made to use SEIR models in forecasting COVID-19 (Giordano et al. (2020), Yang et al. (2020), Bertozzi et al. (2020), Prem et al. (2020), Flaxman et al. (2020)). With the outbreak of COVID-19, accounting for testing has become a critical element in effectively using an SEIR model (Pei, Kandula and Shaman (2020), López and Rodo (2020)).

The MechBayes probabilistic model consists of three parts which together define a probabilistic model for the observed incident cases and deaths with the parameters and state

variables of a compartmental model as latent variables. The core part is the mechanistic disease model $p(x_{1:T}, \eta_{1:T} \mid \theta)$, which defines the distribution of the state variables $x_{1:T}$ and time-varying parameters $\eta_{1:T}$, given a vector θ of fixed, nontime-varying parameters. The state variable x_t is a vector that enumerates the number of individuals in each compartment (susceptible, exposed, infectious, etc.) at time t, while η_t contains parameters of the disease model or observation process that change over time (e.g., due to changes in social distancing or test availability), which we model stochastically. MechBayes operates at a daily time step. The state trajectory $x_{1:T} = (x_1, \ldots, x_T)$ concatenates state vectors from each day, and $\eta_{1:T}$ collects time-varying parameters in a similar fashion. MechBayes also defines a prior distribution $p(\theta)$ over fixed parameters and an observation model $p(y_t \mid x_t, \theta)$ for the vector y_t of observed variables at time t (incident cases and deaths), given the state vector x_t, time-varying parameters η_t, and fixed parameters θ. Each part of the probabilistic model is expressed by writing Python code to sample from the corresponding distribution within the NumPyro probabilstic programming language (Uber (2020)), which automates the construction of Markov chain Monte Carlo algorithms to sample from the distributions $p(\theta, x_{1:T}, \eta_{1:T} \mid y_{1:T})$, for inference about unobserved parameters and state variables, and $p(y_{T+1:T+k} \mid y_{1:T})$, for forecasting (by integrating over unobserved state variables and parameters).

2.3.1. Disease model.—The MechBayes compartmental model is illustrated in Figure 1 and is based on the classical SEIRD framework (Korolev (2021)). It consists of state variables S, E, I, R, D_1, D_2 that indicate the number of individuals in the population that belong on a given day to each one of the following mutually exclusive compartments: susceptible (S), exposed (but not yet infectious) (E), infectious (I), recovered (R), or one of two death compartments (D_1 and D_2).[2] The death pathway is separated into two compartments to incorporate a time-delay between infection and death that is modeled separately from the ratio between observed cases and observed deaths, which both have prior estimates from the literature (Russell et al. (2020)). For simplicity, we assume a closed population of size N. The following parameters govern how members of the population move between compartments:

- β_t: Transmission rate, which we allow to vary by time t;
- σ: Rate of transition from the exposed state E to infectious state I; that is, $1/\sigma$ is the expected duration of the time between exposure and onset of infectiousness;
- γ: Rate of transition from the infectious state I to either D_1 or R; that is, $1/\gamma$ is the expected duration of the infectious period;
- ρ: Fatality rate (i.e., probability of transitioning from I to D_1, as opposed to R);
- λ: Rate of transition from D_1 to D_2 (i.e., the inverse of expected number of days in D_1 compartment before death).

[2]We will also use the state variable name as a short name for the compartment itself–for example, “the E state” –with the correct interpretation always being clear from context.

On a given day t, the following differential equations describe the instantaneuous changes in each compartment with respect to the continuous time index $\tau \in (t,\ t+1]$:

$$\begin{aligned}
\frac{dS}{d\tau} &= -\beta_t \frac{SI}{N}, \\
\frac{dE}{d\tau} &= \beta_t \frac{SI}{N} - \sigma E, \\
\frac{dI}{d\tau} &= \sigma E - \gamma I, \\
\frac{dR}{d\tau} &= (1-\rho)\gamma I, \\
\frac{dD_1}{d\tau} &= \rho\gamma I - \lambda D_1, \\
\frac{dD_2}{d\tau} &= \lambda D_2 .
\end{aligned} \tag{1}$$

In addition, we augment the dynamics with an extra variable $C(\tau)$ to count the cumulative number of individuals that enter the I compartment, with dynamics $\frac{dC}{d\tau} = \sigma E$ that capture only the flow into, and not the flow out of, I. The number of individuals that first become infectious on day t is then $C(t+1) - C(t)$; we consider these individuals candidates for being detected as confirmed cases on day t. We do not attempt to model testing delays or mismatches between onset of infectiousness and onset of a detectable infection.

The state vector at time t is then

$$x_t = (S(t),\ E(t),\ I(t),\ R(t),\ D_1(t),\ D_2(t),\ C(t)).$$

The distribution $p(x_1)$ of the initial state x_1 is described in the Supplementary Material (Gibson, Reich and Sheldon (2023)). While equation (1) is in continuous time, the observed data is in discrete (daily) time. The update from time t to time $t+1$ is obtained by numerically solving the ordinary differential equation (ODE) with dynamics $\frac{dx}{d\tau}$, given by equation (1) for the time interval $\tau \in (t,\ t+1]$, over which the dynamics are fixed. This is contrasted with the typical approach where the dynamics are discretized (say by RK4) to a single step from $[t,\ t+1]$. We are explicitly running the odesolver continuously within the time interval $(t,\ t+1)$ and sampling the right endpoint for a more accurate approximation of the dynamics. We write this as

$$x_{t+1} = \text{odesolve}\left(x_t,\ \frac{dx}{d\tau}\right).$$

We use the ODE solver in the the Python library JAX, which uses the Dormand-Prince algorithm (Dormand and Prince (1980)), a member of the Runge-Kutta family of ODE solvers. Importantly, JAX also supports automatic differentiation of `odesolve` using the adjoint method (Chen et al. (2018)) to compute partial derivatives of x_{t+1} with respect to both the initial value x_t and all dynamics parameters affecting $\frac{dx}{d\tau}$. This is a key functionality that

allows us to embed ODE dynamics within a probabilistic model for which NumPyro can perform inference using Hamiltonian Monte Carlo (Neal (2011)).

In 2020, significant efforts to control the spread of COVID-19 relied on nonpharmaceutical interventions (Catching et al. (2021)). These included mandatory distance between individuals, closures of public spaces, and mask wearing. To add to the complexity, these interventions were implemented and repealed at different time points throughout the year, and the public complied with the interventions to varying degrees (Simonov et al. (2020)). In order to capture the aggregate effect of the interventions and other behavior changes nonparametrically, we choose a flexible model for the time-varying transmission parameter. We allow β_t to vary following a Gaussian random walk on logarithmic scale, that is,

$$\log\ \beta_t \sim \mathcal{N}(\log\ \beta_{t-1},\ \sigma_\beta\ =\ 0.2).$$

The random walk models nonstationary dynamics within the *observed* time period (t from 1 to T). For forecasting ($t > T$), MechBayes does not attempt to model future behavior changes and simply predicts that the final value of β_t will persist in to the future. However, to avoid extreme senstivity of forecasts to one or a few data points near the end of the time series, we average over the last 10 days instead of taking the final value; that is, for all $i \geq 1$,

$$\beta_{T+i} = \frac{1}{10}\sum_{j=0}^{9}\beta_{T-j}.$$

2.3.2. Observation model.—The observed data used to fit the model is based on time series of incident confirmed cases and deaths. The model is fit separately for each location. The observations on day t are $y_t = (y_{t,c},\ \ y_{t,d})$, where $y_{t,c}$ is the number of new cases confirmed on day t and $y_{t,d}$ is the number of new reported deaths. We assume that $y_{t,c}$ is a noisy observation of $C(t+1) - C(t)$, the modeled number of new infections on day t, using the NB2 negative binomial model for for overdispersed counts (Cameron and Trivedi (1986)),

$$y_{t,c} \sim \mathrm{NB2}(\mu_{t,c}, \kappa_c),\ \ \mu_{t,c} = p_{t,c}\cdot(C(t+1) - C(t)).$$

This satisfies $\mathbb{E}[y_{t,c}] = \mu_{t,c}$, with the parameter $p_{t,c}$ acting as a detection rate on the modeled number of new infections; the parameter κ_c controls overdispersion, with $\mathrm{Var}(y_{t,c}) = \mu_{t,c} + \kappa_c\mu_{t,c}^2$. Note that the detection rate $p_{t,c}$ is time varying (see below). Similarly, we assume that $y_{t,d}$ is a noisy observation of $D_2(t+1) - D_2(t)$,

$$y_{t,d} \sim \mathrm{NB2}(\mu_{t,d},\ \ \kappa_d),\ \ \mu_{t,d} = p_d\cdot(D_2(t+1) -\ \ D_2(t)).$$

The detection rate p_d for deaths is not time varying. The dispersion parameters κ_c and κ_d for both cases and deaths are estimated and given a truncated normal prior distribution with location 0.30, scale 0.15, and lower truncation limit 0.10.

We model the time-varying case detection rate as

$$p_{1,c} \sim \text{Beta}(15, 35),$$
$$\text{logit}(p_{t,c}) \sim \mathcal{N}(\text{logit}(p_{t-1,c}), \sigma_{p_c} = 0.2), \quad t \geq 1.$$

The Beta distribution on the case detection rate at time $t = 1$ (corresponding to early March in our operational model) satisfies $\mathbb{E}[p_c] = 0.3$, with 90% probability of falling between 0.22 and 0.38. Preliminary experiments suggested that the detection rate is poorly determined by data, and short-term forecasts are not sensitive to this parameter.[3] We, therefore, use a moderately strong prior centered at 30%, as suggested by the literature (MIDAS (2020)). We then allow the detection rate to vary over time, following a Gaussian random walk on the log-odds scale, as shown above. This is meant to loosely model changes in diagnostic testing over time; in practice, it provides flexibility in the model that likely captures other changes in the relationship between reported cases and deaths over time, such as changes in the fatality ratio of the population infected at a given time.

For deaths we place a strong prior on the reporting rate: $p_d \sim \text{Beta}(90, 10)$. This satisfies $\mathbb{E}[p_d] = 0.9$ with 90% probability between 0.89 and 0.92. That is, we assume that deaths due to COVID-19 are most often correctly reported (Weinberger et al. (2020)). As with the absolute value of the case detection rate, short-term forecasts are not very sensitive to this parameter.

2.3.3. Epidemiological model parameters.—We use informative priors for epidemiological parameters, such as γ, σ, ρ, λ, and initial compartment values based on reasonable estimates of their epidemiological interpretation (Gibson, Reich and Sheldon (2023)). However, these parameters are unidentifiable from the observed data of cases and deaths, and β_t can absorb any misspecification of the true epidemiological parameter settings. While our initial parameter settings were based on external estimates of the incubation period (σ), recovery period (γ), and time until death (λ) (MIDAS (2020)), their exact value is unimportant due to the lack of identifiability.

2.3.4. Implementation.—MechBayes is implemented in the NumPyro probabilistic programming language (Phan, Pradhan and Jankowiak (2019)) which automates the complex task of designing a posterior sampling algorithm. NumPyro uses the JAX Python library (Bradbury et al. (2018)) to automatically compute the partial derivatives needed for sampling via Hamiltonian Monte Carlo (Neal (2011)) and to compile model code for highly efficient computations. JAX includes a differentiable solver for ordinary differential equations (ODEs) (Chen et al. (2018)) which allows us to embed ODE-based compartmental models into the full probabilistic model with relative ease. The components described so far lead to the full probability model

$$p(\theta, \eta_{1:T}, x_{1:T}, y_{1:T}) = p(\theta)p(x_{1:T}, \eta_{1:T} \mid \theta)\prod_{t=1}^{T} p(y_t \mid x_t, \eta_t, \theta),$$

[3]It primarily impacts inferences about the true number of infections in the population; forecasts are, therefore, expected to be more sensitive to this parameter as herd immunity is approached.

where $\eta_t = (\beta_t, \ p_{t,c})$ are time-varying parameters (contact rate and case detection rate) and

$$\theta = [\sigma, \ \gamma, \ \rho, \ \lambda, \ p_d, \ \kappa_c, \ \kappa_d, \ S_1, \ I_1, \ E_1, \ D_{1,1}, \ D_{2,1}, \ R_1]$$

is the vector containing all other parameters.

Each model component is implemented in NumPyro (Phan, Pradhan and Jankowiak (2019)). We then use NumPyro's implementation of the No-U-Turn Sampler (Hoffman and Gelman (2014)) (a variant of Hamiltonian Monte Carlo (Neal (2011))) to draw samples from the posterior distribution $p(\theta, \eta_{1:T}, \ x_{1:T} \mid y_{1:T})$, given an observation sequence $y_{1:T}$ (for model diagnostics), and to sample from the distribution $p(y_{T+1:T+k} \mid y_{1:T})$ to make forecasts of future reported cases and deaths.

We draw 1000 warm-up sample and then 1000 posterior samples of model parameters. We also monitor the number of effective samples produced by HMC to ensure it is large enough to reflect accurate exploration of the posterior (Betancourt (2017)). All $\hat{R}$ values were below 1.2.

2.4. Operational forecasts.

On May 10, 2020, we began submitting incident and cumulative death forecasts on a weekly basis to the U.S. Centers for Disease Control (CDC) through the COVID-19 Forecast Hub consortium (Cramer et al. (2021b)). [4] Each week we submitted one- to four-week ahead forecasts for the 50 U.S. states and Washington, D.C., and later added forecasts for the U.S. national level, and U.S. territories. All forecasts used daily data up to and including the Sunday before the Monday submission. The "one week ahead" forecast corresponds to the week ending on the following Saturday, the "two week ahead" forecast to the week ending on the second following Saturday and so on. The model remained remained stable from May 10, until the time of writing with only minor changes, for example, to prior distributions.

Over time, we developed a quality-assurance process to tune our model and to detect and troubleshoot suspicious forecasts. We regularly monitored the performance of our recent forecasts in terms of mean absolute error and calibration of prediction intervals, as measured by the probability integral transform (Gneiting, Balabdaoui and Raftery (2007)). We used these metrics and diagnostic plots to compare submitted forecasts to alternate models to tune parameters. This led us to introduce a resampling procedure to mitigate too-large prediction intervals (on May 17, 2020) and to slight changes to prior distributions (on September 6 and October 20, 2020).

Suspicious forecast were primarily caused by data reporting issues. It was relatively common for a state to report a large backlog of cases or deaths on one day due to changes in reporting practices or to correct previous errors. As an extreme example, New Jersey (NJ) reported nearly 1600 daily deaths on June 25, 2020, when it began the practice of including deaths from probable COVID-19 cases in its totals. Similarly, Texas (TX) removed 3000

[4]For two weeks prior to May 10, we submitted forecasts of cumulative deaths only while the model was under active development and lacked several of the main components described here.

confirmed cases on July 7, 2020, when it determined that cases detected by antigen testing should not be reported. Changes of smaller magnitude were extremely common. Because MechBayes includes a flexible model of time-varying transmission, it interprets large changes in cases or deaths as evidence of significantly increased or decreased transmission which leads to unrealistic forecasts.

Our quality-assurance process involved viewing diagnostic plots of each forecast together with the recent time series of incident deaths and cases to identify forecasts that were unduly influenced by data reporting issues. We also checked the JHU CSSE website (Dong, Du and Gardner (2020)) for notifications of reporting issues that might not be obvious in diagnostic plots. After identifying a potential problem, we searched for documented evidence of a reporting issue. These were usually reported on state department of health web sites or by local news outlets. If we could identify a reporting issue, we distributed the excess number of incident cases or deaths evenly over some time window selected using our best judgment based on available information.

We made a small number of other interventions. Some states do not report data on Saturdays or Sundays; we modified the data to omit such observations instead of treating them as zeros. Starting in October, we sometimes omitted weekend observations, even if they were nonzero to mitigate the influence of low values that are due only to the weekly reporting cycle. In a small handful of cases, the inference routine failed to converge or diagnostics showed signs of numerical instability; in those cases we adjusted the prior distributions slightly and reran the model to overcome the problem.

2.5. Experimental setup.

We conducted two different evaluations. First, we evaluated the forecasts made in real time and submitted to the CDC via the COVID-19 Forecast Hub to assess the quality of MechBayes as an operational forecast model. Second, we conducted an ablation study that compared retrospective forecasts made using different versions of MechBayes to assess the importance of different model components on forecast accuracy.

2.5.1. Baseline forecast evaluation.—We evaluated all one- to four-week ahead incident death forecasts submitted to the CDC between July 25, 2020, and June 7, 2021, for the 50 U.S. states and Washington, D.C.[5] We computed the absolute error (AE) for point forecast and examined the distributions of absolute errors for different locations, forecast horizons, and dates. We used mean absolute error (MAE) as a summary metric. In addition, to evaluate the uncertainty calibration of our probabilistic forecasts, we measured the empirical coverage rates of the prediction intervals obtained from the forecasted quantiles by measuring the fraction of actual observed values that fell within different prediction intervals, referred to as coverage. We compared the performance of MechBayes to the performance of the COVID-19 Forecast Hub baseline model described by Cramer et al. (2021a).

[5]We submitted forecasts for U.S. territories and the U.S. as a whole, starting after May 10, but omit these from evaluation to allow for the largest number of evaluation dates where forecasts were made across a consistent set of regions.

2.5.2. Forecast hub alternate model comparison.—To evaluate the relative performance of MechBayes against other models submitted to the forecast hub, we chose the nine models (including MechBayes) that have been submitting forecasts from July 25, 2020, to June 7, 2021, for incident deaths across all 50 states and D.C. for every forecast week. These criteria balance including as many models as possible, including ones that have performed well in other analyses, while also having a large number of locations and dates for which all of the models made forecasts. For each of the models, we examined the distribution of absolute error of all point forecasts as well as summary metrics, such as the mean and median absolute error. We include this analysis to demonstrate that for a particular common subset of locations and dates, MechBayes is a top performing model.

A comprehensive evaluation of forecast hub models by Cramer et al. (2021a) examines multiple performance metrics and addresses the problem of comparing models that make forecasts for different sets of locations and dates and also finds that MechBayes is the top component model across many different evaluation criterion and the largest subset of common locations and dates (all 50 states from July 25, 2020, through June 7, 2021). This evaluation allows us to claim that MechBayes is the best performing component COVID-19 mortality forecasting model submitted to the CDC challenge out of 26 nonensemble models evaluated.

2.5.3. Ablation test.—We also performed a retrospective evaluation to demonstrate the improvement in accuracy due to addition of different model components. We define the following three variants of MechBayes:

- MechBayes Full is the full MechBayes model as submitted to the forecast hub and described in the previous sections, including observations of both reported cases and deaths and a time-varying random-walk model for the case detection rate $p_{t,c}$.
- MechBayes Fixed-Detection is the same as MechBayes full but with the time-varying detection rate $p_{t,c}$ replaced by a constant detection rate p_c.
- MechBayes Death-Only is the same as MechBayes Fixed-Detection but with observations only on incident deaths (the forecasted quantity) and not on cases. This model is included to assess the value of using incident cases as evidence to help forecast incident deaths.

Other than the changes described above, all model components, data handling, and fitting procedures were identical. Note that we did not include a model without time-varying transmissibility. Such a model is unable to adequately fit the observed data; previous COVID-19 modeling has clearly established that time-varying transmissibility is an essential model component (Pei, Kandula and Shaman (2020), Smirnova, deCamp and Chowell (2019), Flaxman et al. (2020), Abbott et al. (2020)).

3. Results.

3.1. Model fitting and inference.

MechBayes captures signal in the observed data, even in the presence of highly variable incident death reporting, and produces forecasts and prediction intervals that track the data well (Figure 2(A)). The model infers a relationship between the logarithm of incident deaths and time that is nearly linear over short time periods but with slopes that not only change over time at somewhat discrete time points (Figure 2(B)), highlighting the exponential growth (or decline) over short time periods that is a hallmark of compartmental models, but also the fact that these dynamics vary over longer time periods.

The inferred value of the time-varying contact rate parameter β_t (Figure 2(C)) is closely tied to the observed rate of change of incident deaths (and cases, not shown), especially as observed on a logarithmic scale: β_t is high across all four example states in mid-March, when incident deaths grew rapidly, then falls as the growth rate of incident deaths declines during and after the initial wave, with subsequent changes that can be matched to specific events in the states, for example, increases in β_t associated with second waves in Texas, California, and Florida during the summer of 2020, and a slow increase in β_t in New York associated with an eventual increase in deaths in the fall of 2020. In all four states, the inferred value of time-varying case detection rate $p_{t,c}$ increases significantly from the start of the pandemic (Figure 2(D)). In practice, this parameter likely functions to model *any* changes over time in the ratio of observed cases to observed deaths. One reason for such a change is increased diagnostic testing; another reason is a decrease in the overall infection-fatality ratio (e.g., due to changes in the age distribution of patients and improved treatments). Both changes would lead to a larger number of observed cases for the same number of deaths and likely occurred in conjunction, leading the model to significantly increase its estimate of $p_{t,c}$ over time. It is apparent that MechBayes also uses $p_{t,c}$ to absorb some reporting anomalies, as seen in Texas: a string of both abnormally high and low numbers of incident deaths were reported in late summer, which correspond to the model inferring a temporary decrease in $p_{t,c}$.

3.2. Real-time forecast results.

3.2.1. Baseline comparison.—MechBayes had lower absolute error than the baseline model in all quantiles of the error distribution (Figure 4(A)). The gap in performance (as measured by absolute error) increased as the magnitude of the error increased for all quantiles of the error distribution. MechBayes also had a lower absolute error at the central tendency of the absolute error distribution (as measured by mean or median) (Table 1).

Overall, MechBayes had an MAE of 50.00 deaths, when averaging over all regions, forecast dates, and targets. The baseline model had an MAE of 76.62 deaths. The prediction intervals at the 95% level covered the truth 94.8% of the time for MechBayes, compared to 89.2% for the baseline model over all targets, regions, and forecast dates (Figure 3(B)).

MechBayes had similar or lower MAE than the baseline for almost all states and targets (Figures 3(A), 3(C)). In particular, for the locations with the highest total death counts (NY,

TX, CA, FL) (Dong, Du and Gardner (2020)), MechBayes uniformly outperformed the base-line.

MechBayes improved uniformly over the baseline model for every target (Figure 3(C)). The mean absolute error increased as horizon increased which is to be expected. The distribution of errors by forecast date for a given target showed significant variability, suggesting that different targets were easier to predict on certain forecast dates, again reflecting the change in forecast difficulty throughout epidemic.

MechBayes had lower MAE by forecast date (averaged over locations and targets) than the baseline for 40 out of 46 forecast weeks (Figure 4(B)). The largest increase in incident deaths during the evaluation period occurred in winter 2020. MechBayes significantly outperformed the baseline model during these weeks, with the exception of forecasts made on December 26, 2020, and January 2, 2021. However, in weeks with a small increase or a decrease in incident deaths the MAEs were much closer. This suggests again that MechBayes performs well during periods of more rapid change in incident deaths.

MechBayes prediction intervals contained the truth with at least the predicted probability (Figure 3(D)) but were somewhat conservative: the empirical probability of containing the truth was nearly exact for the 95% interval and higher than predicted for smaller intervals.

MechBayes predictive distributions also tended to be wider than the baseline model (Figure 4 D) for all levels of theoretical coverage. The wide prediction intervals, coupled with the lower absolute error of MechBayes relative to the baseline (Table 1), suggests that MechBayes predictive distributions are both closer to the true value and have wider tails than the baseline model. At lower levels of theoretical coverage, the wider tails are underconfident. However, above 0.8 theoretical coverage levels MechBayes is well calibrated.

3.2.2. Alternate model comparison.—MechBayes had a lower absolute error in nearly all quantiles of the error distribution for seven out of the eight alternate models submitted to the forecast hub (Figure 4(C)). MechBayes was in the top two out of the nine models based on mean, median, and the 0.95 quantile of the absolute error distribution (Table 1). We subset to models that have been submitting consistently over the course of the pandemic to avoid the complexities of evaluation in the presence of missing forecasts handled in Cramer et al. (2021a). MechBayes also performed better than all models, except the ensemble for most quantiles of the error distributions (Figure 3(A)), with the only exception of the Karlen-pypm model for quantiles between 400 and 600 absolute error units.

The median of the error distribution for all models was between 17 and 34 deaths (Table 1). The mean absolute error was between 49 and 77 deaths. The 0.95 quantile of the error distribution began to show more significant separation between models (Table 1).

3.3. Ablation test results.

MechBayes produced consistently better point forecasts than MechBayes Death-Only or MechBayes Fixed-Detection (Figure 5(A)). When averaged over all targets, locations, and

forecast dates, MechBayes had an MAE of 48.67, MechBayes Death-Only had an MAE of 68.33, and MechBayes Fixed-Detection had an MAE of 79.40. At every quantile level, the error of MechBayes Fixed-Detection was significantly larger than that of MechBayes Death-Only which suggests that using deaths as evidence is only beneficial in conjunction with flexibility allowed by the time-varying ratio between cases and deaths ($p_{t,c}$). Both MechBayes Fixed Detection and MechBayes Death-Only were less well calibrated than MechBayes (Figure 5(B)). The prediction intervals for both models were too narrow at all levels of coverage.

4. Discussion.

MechBayes is a fully Bayesian compartmental model capable of accounting for varying transmission rates, observations on both cases and deaths, and a time-varying ratio of cases to deaths. MechBayes produced consistently accurate real-time forecasts over the course of 23 evaluation weeks, and was ranked as one of the top two of 10 models on median and mean absolute error. Our experiments led us to the following conclusions about the performance of this model and the underlying methodology:

- *MechBayes is accurate when compared to a baseline model*. MechBayes had lower absolute error than the baseline model in almost all quantiles of the error distribution (Figure 4). The performance gain was higher when predicting deaths was difficult (in the upper quantiles of the absolute error distribution) because deaths were changing rapidly. This is true across target, week, and region breakdown. Error is significantly lower for one- to four-week ahead predictions with larger improvements at longer horizons. Additionally, the biggest gains in performance occur in regions with the largest incident death counts (Figure 3(A)), such as Texas (TX), California (CA), New York (NY), New Jersey (NJ), and Florida (FL) (Dong, Du and Gardner (2020)). Finally, MechBayes performance gain was highest in forecast weeks with the large absolute error (Figure 3(B)). This leads us to conclude that MechBayes is better than the baseline model when it really counts—in regions where deaths were high and in weeks that were difficult to predict because of rapidly changing incident deaths.
- *MechBayes is accurate when compared to the alternate models submitted to the forecast hub*. MechBayes ranked second out of nine models in terms of mean and median of the absolute error distribution (Table 1). MechBayes trailed the COVID Hub ensemble by a small margin, even though MechBayes itself was a component model of the ensemble. In an independent evaluation by Cramer et al., MechBayes was the best component model on a variety of probabilistic scoring measures, suggesting it is not only a good point forecasting model (as measured by MAE) but also a good probabilistic forecasting model (Cramer et al. (2021a)). However, the same mechanisms—the underlying exponential growth intrinsic to compartmental models and the flexible, time-varying transmission—that allow MechBayes to accurately model the pandemic in many situations also make its forecasts highly sensitive to errors estimating the current rate of exponential growth. For example, the four-week forecast

for Florida (FL) on July 25, 2020, was too high by 2861 deaths, due to MechBayes estimating a high exponential growth based on recent trends and possible reporting issues, when the eventual growth rate over the next four weeks was much more modest.

- *MechBayes is probabilistically well calibrated*. The MechBayes 95% prediction interval contains the truth 94.6% of the time (Figure 3(D)). MechBayes is conservative for smaller intervals. As a Bayesian model, MechBayes is able to reason effectively about uncertainty in the epidemiological model parameters, state variables, and observation noise, given the evidence, and translate this into appropriately calibrated forecast uncertainty.
- *Adding case data when predicting deaths is helpful but only when accounting for time-varying relationship between observed cases and deaths.* Allowing for a time-varying ratio between cases and deaths is a key feature for lower MAE (Figure 5(A)). MechBayes FULL both incorporates incident cases as evidence and allows for a flexible deviation between cases and deaths, which makes the model consistently more accurate than a model that does not account for cases at all (MechBayes Death-Only) and a model that does account for cases but fixes the detection probability (MechBayes Fixed-Detection). Including cases without properly accounting for factors that yield a changing ratio between observed cases and observed deaths over time hurts performance over leaving out observations on cases all together.

We have seen that MechBayes is a powerful Bayesian compartmental model that can capture the real-world complexities of forecasting during a pandemic. MechBayes' disease model is a classical compartmental model which has good inductive bias for a novel epidemic. MechBayes is fully Bayesian which allows for a balance between model structure, evidence through observations on cases and deaths, and uncertainty. The implementation in the NumPyro probabilistic programming language allowed for rapid model development and experimentation. Finally, a reasonable investment of effort in validation prevented model pathologies due to data quality issues.

While we chose an exponential random walk on β_t, there are many other choices for flexible nonparametric modeling of transmissibility. Further work might consider a spline model, Gaussian process, or a semiparametric model capable of taking intervention dates as covariates. The reproductive rate of COVID-19 has been fluctuating around 1 for the duration of the pandemic (Abbott et al. (2020)). A model that smooths transmissibility toward a reproductive ratio of 1 may improve forecast skill by avoiding large growth or decline. Further work might explore more structured models on β_t that may improve forecast skill. While there are many potential models for β_t, there is a trade-off between complexity and calibration. Increasing flexibility may inflate prediction intervals. Additionally, as more COVID-19 data streams come online, more observation models on compartments can be added to MechBayes and fit using the same framework. Other methods of expressing compartmental models (e.g., the renewal framework in Flaxman et al. (2020)) may lead to more efficient and flexible implementations. Modeling more characteristics of the surveillance system (such as weekly reporting) may also improve forecast performance.

Through real-time and retrospective evaluation, we demonstrated the success of MechBayes in forecasting COVID-19, both in terms of point and probabilistic forecasts. The model is able to improve over the baseline model as well as reduced forms of MechBayes and is ranked in the top two models out of the nine considered that submitted forecasts to the forecast hub for a year's worth of evaluation dates. While future pandemics may be unavoidable, MechBayes is a flexible framework that can be adapted to the unique challenges of pandemic forecasting efforts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

We would like to thank Evan Ray for many productive conversations during the development of MechBayes. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIGMS or the National Institutes of Health.

Funding.

This material is based upon work supported by the National Science Foundation under Grant No. 1749854. This work has been supported by the National Institutes of General Medical Sciences (R35GM119582).

REFERENCES

Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, Munday JD, Meakin S, Doughty EL et al. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. 5 112.

Bertozzi AL, Franco E, Mohler G, Short MB and Sledge D (2020). The challenges of modeling and forecasting the spread of COVID-19. Proc. Natl. Acad. Sci. USA 117 16732–16738. MR4242725 10.1073/pnas.2006520117 [PubMed: 32616574]

Betancourt M (2017). A conceptual introduction to Hamiltonian Monte Carlo. Preprint. Available at arXiv:1701.02434.

Bokler B (1993). Chaos and complexity in measles models: A comparative numerical study. Math. Med. Biol 10 83–95.

Borchering RK, Viboud C, Howerton E, Smith CP, Truelove S, Runge MC, Reich NG, Contamin L, Levander J et al. (2021). Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—United States, April–September 2021. Morb. Mort. Wkly. Rep 70 719.

Bradbury J, Frostig R, Hawkins P, Leary C, Maclaurin D, Necula G, Paszke A, Vanderplas J, Wanderman-Milne S and Zhang Q (2018). JAX: Composable transformations of Python+NumPy programs. Available at http://github.com/google/jax.

Cameron CA and Trivedi PK (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. J. Appl. Econometrics 1 29–53.

Catching A, Capponi S, Yeh MT, Bianco S and Andino R (2021). Examining the interplay between face mask usage, asymptomatic transmission, and social distancing on the spread of COVID-19. Sci. Rep. 11 1–11. [PubMed: 33414495]

Chen RT, Rubanova Y, Bettencourt J and Duvenaud DK (2018). Neural ordinary differential equations. In Advances in Neural Information Processing Systems 6571–6583.

Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Rivadeneira AJC, Gerding A, Gneiting T, HOUSE KH et al. (2021a). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. medRxiv.

Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, Brennen A, Rivadeneira AJC, Gerding A et al. (2021b). The United States COVID-19 forecast hub dataset. medRxiv.

Dong E, DU H and Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 533–534. [PubMed: 32087114]

Dormand JR and Prince PJ (1980). A family of embedded Runge–Kutta formulae. J. Comput. Appl. Math. 6 19–26. MR0568599 10.1016/0771-050X(80)90013-3

Dukic V, Lopes HF and Polson NG (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. J. Amer. Statist. Assoc. 107 1410–1426. MR3036404 10.1080/01621459.2012.713876

Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, Whittaker C, Zhu H, Berah T et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 584 257–261. [PubMed: 32512579]

Frasso G and Lambert P (2016). Bayesian inference in an extended SEIR model with nonparametric disease transmission rate: An application to the Ebola epidemic in Sierra Leone. Biostatistics 17 779–792. MR3604280 10.1093/biostatistics/kxw027 [PubMed: 27324411]

Gibson GC, Reich NG and Sheldon D (2023). Supplement to “Real-time mechanistic Bayesian forecasts of COVID-19 mortality.” 10.1214/22-AOAS1671SUPPA, 10.1214/22-AOAS1671SUPPB

Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A and Colaneri M (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat. Med. 26 855–860. [PubMed: 32322102]

Gneiting T, Balabdaoui F and Raftery AE (2007). Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B. Stat. Methodol 69 243–268. MR2325275 10.1111/j.1467-9868.2007.00587.x

Grinsztajn L, Semenova E, Margossian CC and RIOU J (2021). Bayesian workflow for disease transmission modeling in Stan. Stat. Med 40 6209–6234. MR4339396 10.1002/sim.9164 [PubMed: 34494686]

Hall I, Gani R, Hughes H and Leach S (2007). Real-time epidemic forecasting for pandemic influenza. Epidemiol. Infect. 135 372–385. [PubMed: 16928287]

Hoffman MD and Gelman A (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res 15 1593–1623. MR3214779

Hotta LK (2010). Bayesian melding estimation of a stochastic SEIR model. Math. Popul. Stud 17 101–111. MR2664909 10.1080/08898481003689528

Johndrow J, Ball P, Gargiulo M and Lum K (2020). Estimating the number of SARS-CoV-2 infections and the impact of mitigation policies in the United States. Harv. Data Sci. Rev..

Karlen D (2020). Characterizing the spread of COVID-19. Preprint. Available at arXiv:2007.07156.

Kermack WO and Mck AG (1927). A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci. 115 700–721.

KOROLEV I (2021). Identification and estimation of the SEIRD epidemic model for COVID-19. J. Econometrics 220 63–85. MR4185125 10.1016/j.jeconom.2020.07.038

Krantz SG and RAO ASS (2020). Level of under-reporting including under-diagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. Infect. Control Hosp. Epidemiol 41 857–859. [PubMed: 32268929]

Lau H, Khosrawipour T, Kocbach P, Ichii H, Bania J and Khosrawipour V (2021). Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. Pulmonology 27 110–115. 10.1016/j.pulmoe.2020.05.015 [PubMed: 32540223]

Lekone PE and Finkenstädt BF (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. Biometrics 62 1170–1177. MR2307442 10.1111/j.1541-0420.2006.00609.x [PubMed: 17156292]

López L and Rodo X (2020). A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics. Available at SSRN 3576802.

Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, Fernansez D, Greene SK, Kipshidze N et al. (2019). Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. BMC Public Health 19 1659. [PubMed: 31823751]

Mbuvha R and Marwala T (2020). Bayesian inference of COVID-19 spreading rates in South Africa. PLoS ONE 15 e0237126. 10.1371/journal.pone.0237126vaditho [PubMed: 32756608]

Midas (2020). COVID-19 parameter estimates. Available at https://github.com/midas-network/COVID-19/tree/master/parameter_estimates/2019_novel_coronavirus.

Myers MF, Rogers D, COX J, Flahault A and Hay SI (2000). Forecasting disease risk for increased epidemic preparedness in public health. Adv. Parasitol 47 309–330. [PubMed: 10997211]

Neal RM (2011). MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods 113–162. CRC Press, Boca Raton, FL. MR2858447

Ong JBS, Mark I, Chen C, Cook AR, Lee HC, Lee VJ, Lin RTP, Tambyah PA and Goh LG (2010). Real-time epidemic monitoring and forecasting of H1N1–2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PLoS ONE 5 e10036. [PubMed: 20418945]

Osthus D, Hickmann KS, Caragea PC, Higdon D and Del Valle SY (2017). Forecasting seasonal influenza with a state-space SIR model. Ann. Appl. Stat 11 202–224. MR3634321 10.1214/16-AOAS1000 [PubMed: 28979611]

Pei S, Kandula S and Shaman J (2020). Differential effects of intervention timing on COVID-19 spread in the United States. medRxiv. 10.1101/2020.05.15.20103655

Phan D, Pradhan N and Jankowiak M (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. Preprint. Available at arXiv:1912.11554.

Prem K, Liu Y, Russell TW, Kucharski AJ, EGGO RM, DAVIES N, FLASCHE S, CLIFFORD S, PEARSON CA et al. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. Lancet Public Health 5 E261–E270. [PubMed: 32220655]

Rahmandad H, Lim TY and Sterman J (2020). Estimating COVID-19 under-reporting across 86 nations: Implications for projections and control. Available at SSRN 3635047.

Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, Bracher J, Zheng A, Yamana TK et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. medRxiv.

Russel TW, Hellewell J, Abbot S et al. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. Available at the Centre for Mathematical Modelling of Infectious Diseases Repository.

Russell TW, Hellewell J, Jarvis CI, Van Zansvoort K, Abbott S, Ratnayake R, Flasche S, Eggo RM, Edmunds WJ et al. (2020). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. Euro Surveill. 25 2000256. [PubMed: 32234121]

Shaman J and Karspeck A (2012). Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad. Sci. USA 109 20425–20430. [PubMed: 23184969]

Simonov A, Sacher SK, Dubé J-PH and BISWAS S (2020). The persuasive effect of fox news: Non-compliance with social distancing during the COVID-19 pandemic. Technical report, National Bureau of Economic Research.

Smirnova A, Decamp L and Chowell G (2019). Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. Bull. Math. Biol 81 4343–4365. MR4034830 10.1007/s11538-017-0284-3 [PubMed: 28466232]

Syafruddun S and Noorani M (2012). SEIR model for transmission of dengue fever in Selangor Malaysia. Int. J. Mod. Phys. Conf. Ser 9 380–389.

Uber Labs Ai (2020). “NumPyro.” Available at https://readthedocs.org/projects/numpyro/downloads/pdf/stable/.

Weinberger DM, Chen J, Cohen T, Crawford FW, Mostashari F, Olson D, Pitzer VE, Reich NG, Russi M et al. (2020). Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. JAMA Intern. Med 180 1336–1344. [PubMed: 32609310]

Yang H and Lee J (2020). Variational Bayes method for ODE parameter estimation with application to time-varying SIR model for COVID-19 epidemic. Preprint. Available at arXiv:2011.09718.

Yang Z, Zeng Z, Wang K, Wong S-S, Liang W, Zanin M, Liu P, Cao X, Gao Z et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis 12 165. [PubMed: 32274081]

Zhuang L and Cressie N (2014). Bayesian hierarchical statistical SIRS models. Stat. Methods Appl 23 601–646. MR3278930 10.1007/s10260-014-0280-9

Zhuang L, Cressie N, Pomeroy L and Janies D (2013). Multi-species SIR models from a dynamical Bayesian perspective. Theor. Ecol 6 457–473.

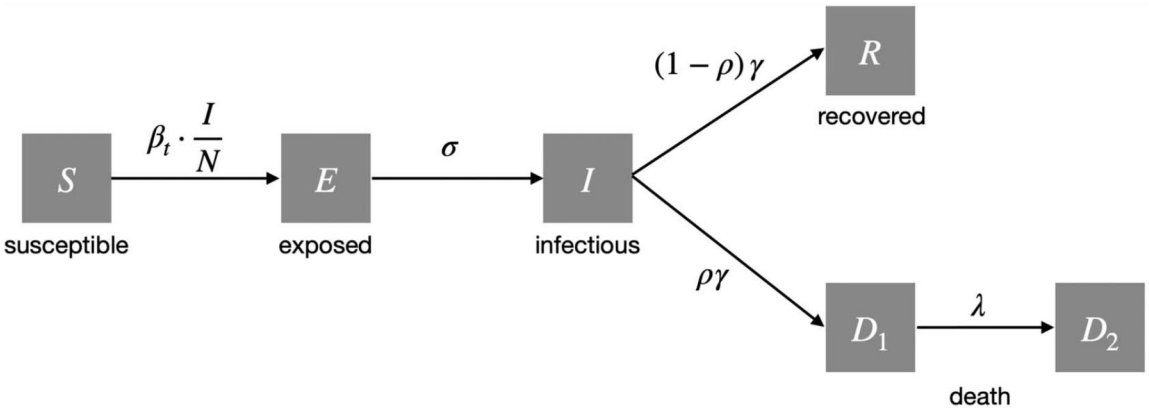

Fig. 1.

Flow diagram for MechBayes. Susceptibles (S) become exposed (E) with a rate of $\beta_t \cdot \frac{I}{N}$ (proportional to the number of infected and infection probability times average number of contacts). Exposed individuals become infectious with a mean time of $\frac{1}{\sigma}$. Infectious individuals can either recover or enter a D_1 compartment, representing individuals who will eventually succumb to the disease, with probability ρ and after a mean time of $\frac{1}{\gamma}$. Individuals in D_1 then enter the final death compartment D_2 with mean time $\frac{1}{\lambda}$. The distinction between D_1 and D_2 aids in accounting and helps separate out a parameter governing the time between infectiousness and death which is useful for model parameterization.

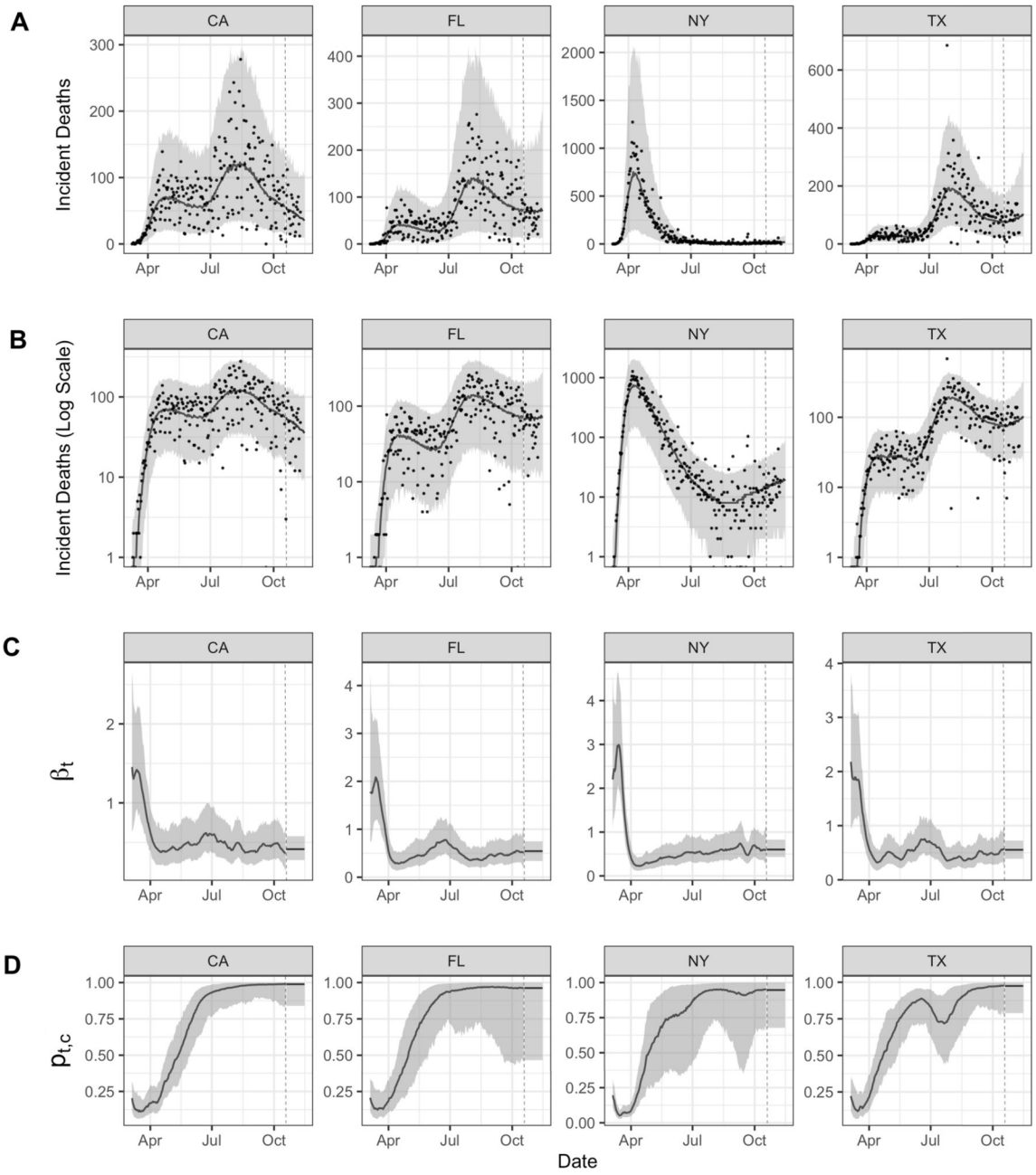

Fig. 2.

A, B. *Example posterior fits as well as one- to four-week ahead forecasts made on October* 18, 2020, *for four selected states. Shaded regions show* 95*% prediction intervals for in-sample* (*before dashed vertical line*) *and forecast* (*after dashed vertical line*) *posterior predictive distributions*; *lines show posterior medians*; *points show observed data*. C. *Posterior median and* 95*% credible interval of time-varying contact rate* β_t *for each of the four states*. D. *Posterior median and* 95*% credible interval of the time-varying ratio between cases and deaths parameter* ($p_{t,c}$).

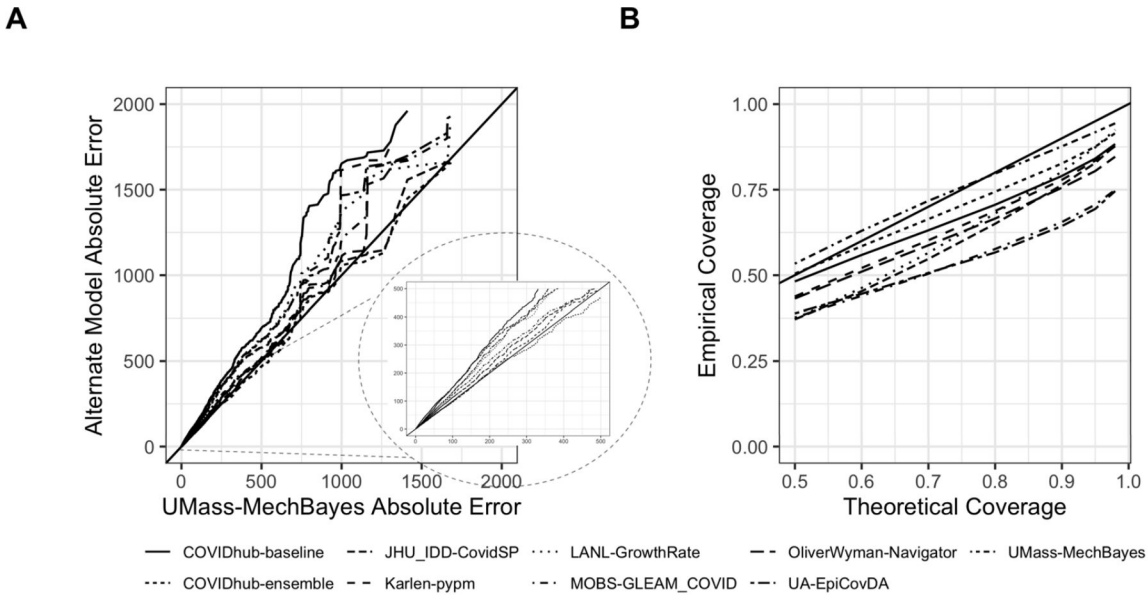

Fig. 3.

A. *Quantile–quantile plot of absolute error distribution for MechBayes* (*x-axis*) *vs. alternate models* (*y-axis*) *submitted to the forecast hub. Each point represents the absolute error for a combination of location, forecast date, and target for July*, 25 2020 *through June* 6, 2021. *MechBayes is consistently outperforming the baseline and alternate models for almost all quantiles of the error distribution.* B. *Coverage probability at the* 50 *%*, 60 *%*, 70 *%*, 80 *%*, 90 *%*, 95 *%, and* 98 *% levels for MechBayes and the alternate models. MechBayes intervals are slightly too wide at the lower theoretical coverage values but are well calibrated at the upper levels and outperform all other models considered.*

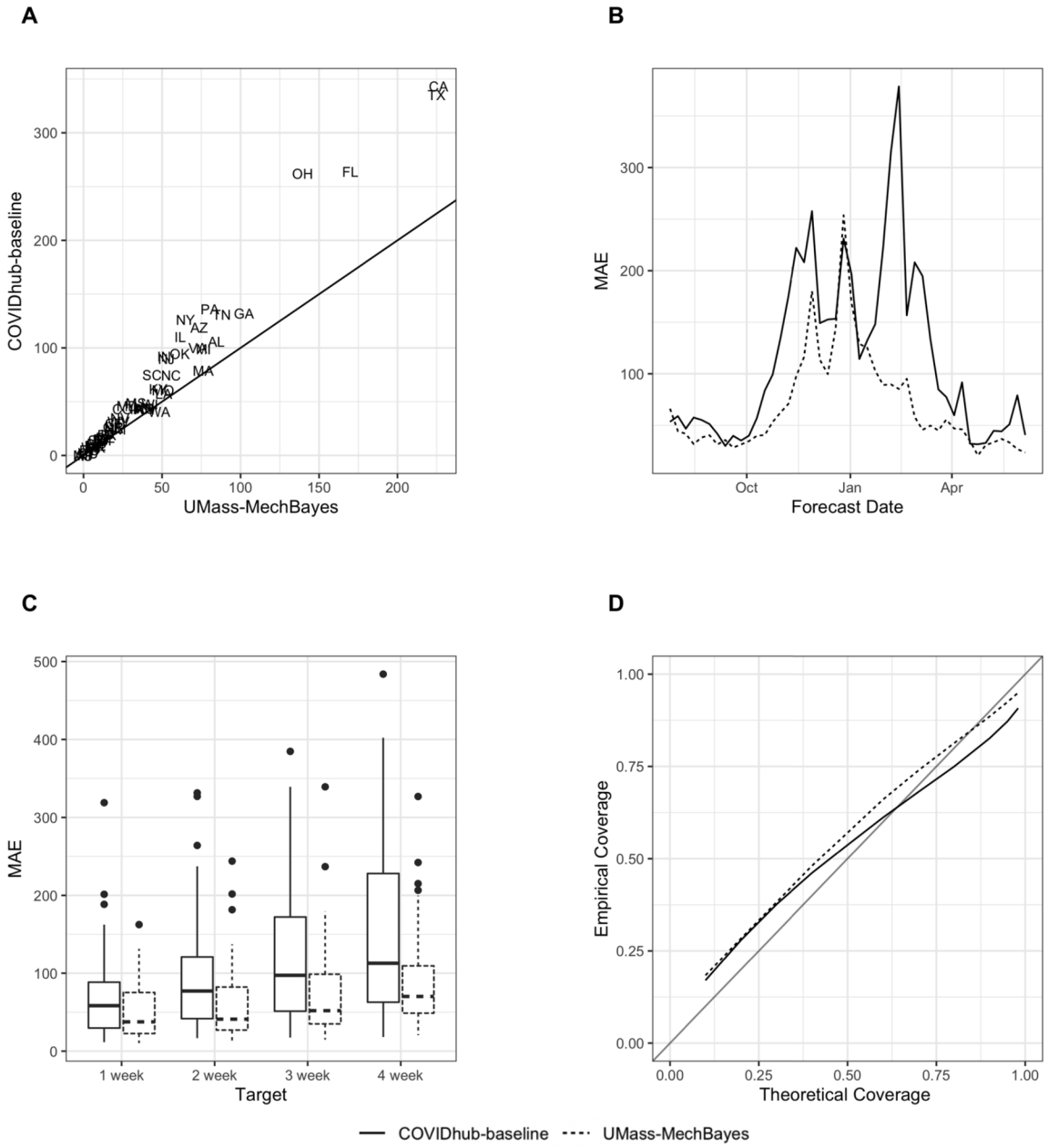

Fig. 4.

A. *Mean absolute errors for MechBayes and the baseline model averaged over all forecast dates and targets for each location. Notice that for states with the largest number of deaths, New Jersey* (*NJ*), *New York* (*NY*), *Florida* (*FL*), *Texas* (*TX*), *California* (*CA*), *MechBayes uniformly outperforms the baseline*. B. *Mean absolute errors for MechBayes and the baseline model averaged over all regions and targets by target end date: a point on panel B represents the absolute error of the one- to four-week ahead forecast made for that date. MechBayes has lower mean absolute error for* 40 *of the* 46 *forecast dates*. C. *Mean absolute error box plots for MechBayes and baseline model by target. Each box plot shows the distribution of MAE values for all forecast dates, where one data point is the MAE over all locations for a single date. MechBayes has lower quartiles of mean*

absolute error across all targets. D. *Prediction interval empirical coverage for MechBayes and the COVIDhub-baseline model for* 0.1–0.98 *levels of theoretical coverage*. *MechBayes has wide intervals at the levels of coverage below* 0.8 *but is well calibrated at higher levels of coverage*.

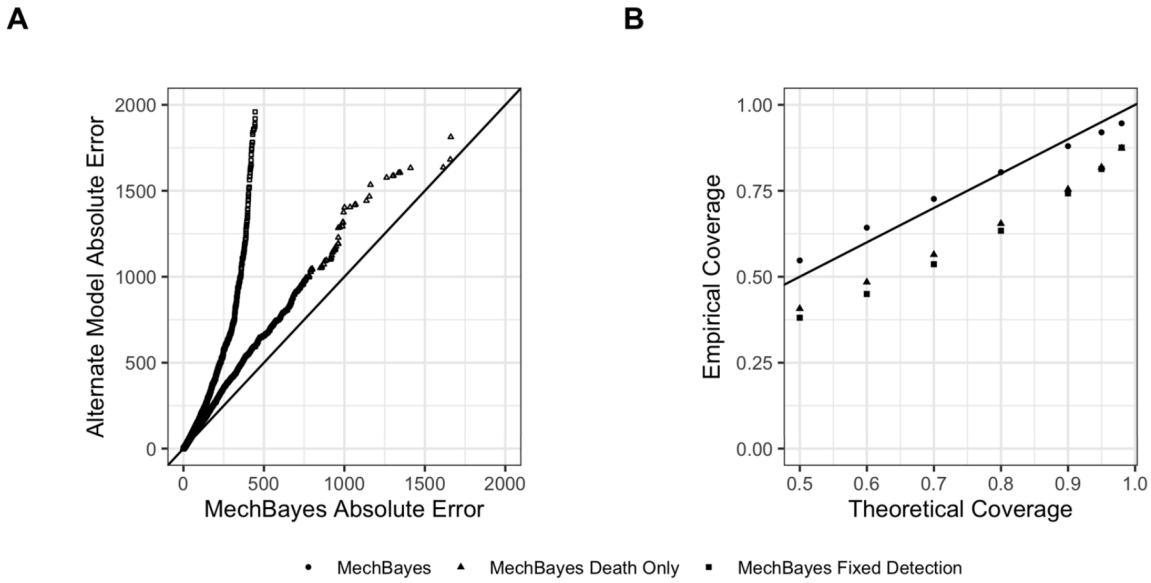

Fig. 5.

A. *Absolute error quantiles of MechBayes* (*y-axis*) *against the reduced models, MechBayes Death-Only and* MechBayes Fixed-Detection. MechBayes *uniformly improves over* MechBayes Fixed-Detection *and improves in all but the maximum quantile over* MechBayes Death-ONLY. B. *Percent of observations* (*y-axis*) *falling within the prediction interval at the given confidence level* (*x-axis*). MechBayes Fixed Detection *seems to be closest to the nominal level of coverage, suggesting that adding the uncertainty in the ratio between observed cases and observed deaths made the model slightly underconfident. In contrast, using only observations on deaths significantly compromised model uncertainty.*

Table 1

Mean, median, and 95% quantiles of the absolute error distribution for state-week-target specific absolute errors for alternate models submitted to the CDC forecast challenge for forecast dates July 25, 2020, to June 6, 2021, across all 50 states. MechBayes has comparable performance to the ensemble model even though MechBayes itself is a component model of the ensemble

	Mean AE	Median AE	95% Quantile of AE
UMass-MechBayes	50.00	17	[0.00, 326.12]
COVID Hub-Baseline	76.62	24	[0.00, 492.12]
COVID Hub Ensemble	49.64	17	[0.00, 308.12]
UMass-MechBayes	51.00	17	[0.00, 326.12]
LANL-GrowthRate	67.67	23.88	[0.33, 438.03]
JHU_IDD-CovidSP	76.30	33.79	[0.97, 457.62]
MOBS-GLEAM_COVID	61.51	22.15	[0.411, 394.48]
UA-EpiCovDA	67.21	23	[1.00, 452.12]
Karlen-pypm	55.67	18.1	[0.40, 344.62]