Contents

MECHANISTIC BAYESIAN FORECASTS OF COVID19

GRAHAM C. GIBSON, NICHOLAS G. REICH, DANIEL SHELDON

Abstract

The COVID-19 pandemic emerged in late December 2019. In the first six months of the global outbreak, the US 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, data quality suffers from under-reporting, delayed reporting due to reporting infrastructure issues, and limited testing. Second, interventions are time-varying across different geographies leading to large changes in transmissibility over the course of the pandemic. Finally, the unidentifiability of the epidemiological parameters leads to difficulty in model fitting. We propose a novel Bayesian compartmental model (MechBayes) that builds upon the classic compartmental framework of susceptible-exposed-infected-recovered (SEIR) model to operationalize COVID-19 forecasting in real time. This includes non-parametric modeling of varying transmission rates, non-parametric modeling of case and death discrepancies due to testing and reporting issues, and finally joint observation likelihood on new case counts and new deaths. The model has been used to submit forecasts to the US Centers for Disease Control, through the COVID-19 ForecastHub [1]. We examine the performance relative to a baseline model in addition to performing an ablation test of our extensions to the classic SEIR models. We demonstrate a significant gain in both point and probabilistic forecast scoring measures using MechBayes.

1. Introduction

The emergence of COVID-19 in early 2020 in the United States led to the largest pandemic in over a century. Understanding the future trajectory of the impact in terms of healthcare burden, economic, treatments and public health response. Forecasts of incident and cumulative deaths due to COVID-19 help in resource allocation, vaccine clinical trial planning, and re-opening strategies [2]. Forecasts provide important data to decision-makers and the general public and can improve situational awareness of current trends and how they will continue in coming weeks.

Infectious disease forecasting, at the time horizon of up to 4 weeks in the future, has benefited public health decision makers during annual influenza outbreaks [3, 4]. However, many forecasts of seasonal disease, such as influenza, often rely on ample historical data to look for patterns that can be projected forward into the future. In an emerging pandemic situation, models must be able to fit to limited data.

In modeling the COVID-19 pandemic, many research groups have turned to the use of compartmental models to explain the underlying transmission of a disease through a population. First introduced by Kermack and McKendrick, such models assume the each individual is in one of a mutually exclusive set of compartments, typically either the susceptible, exposed, infected, or recovered compartment [5]. 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 pandemic represents a unique opportunity to explore their forecasting properties in real-time at both local and global scales.

Our main goal in developing MechBayes is to forecast observed incident deaths, and because we have limited historical data on COVID-19, we think compartmental models are among the most parsimonious models available for forecasting. Our focus is not on inference but forecasting. Therefore, identifying internal parameters of the model, many of

which are poorly determined or not identifiable from the available data. We distinguish this from 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. These models, however, are often not flexible enough for real-time forecasting [**?**].

We introduce a novel operational forecast model based on mechanistic foundations and tailored to the particular needs and data availability of COVID-19. These include, but are not limited to, severe under reporting of cases due to low testing rates especially in mild or asymptomatic cases, time-varying testing rates, delayed reporting, and both the addition and removal of control measures such as social distancing, lockdown, and mask use. We accomplish this by modeling transmission and deviations from cases and deaths (beyond the case fatality ratio) non-parametrically. We choose a Bayesian framework that allows for uncertainty in the epidemiological parameters that are unidentifiable from the data, introducing flexibility suited to forecasting. Implementation in a Bayesian framework also allows for use of a cutting edge probabilistic programming framework for computational speed.

We demonstrate the success of the model in both forecast submissions to the US Centers for Disease Control via the COVID-19 Forecast Hub as well as an ablation model comparison to demonstrate the additional forecast accuracy of our extensions beyond the basic SEIR model. In what follows we first describe the available data and forecast submission infrastructure, outline the basic SEIR compartment model, describe our extensions for real-world pandemic forecasting, and finally evaluate the model using both real-time evaluation from submissions and a retrospective model component analysis.

2. Related Work

Compartmental models have been used to effectively model and forecast disease in non-pandemic situations both retrospectively and in real-time. These include complex compartmental models for real-time influenza forecasting [6][7][8], and a retrospective model evaluation of the 1918 influenza pandemic [9]. Compartmental models have been used not just in respiratory disease but in Ebola [10], measles [11], dengue [12] and a wide variety of other diseases

Compartmental models have also been adopted into a Bayesian framework before, including both stochastic disease dynamics and deterministic dynamics [13][14]. Non-parametric transmissibility was included in a Bayesian SEIR model to study Ebola by Frasso and Lambert [15]. Time-varying transmissibility has also been studied in the frequentist setting using complex non-parametric functions [16]. Many efforts have been made to use SEIR models in forecasting COVID-19 [17][18] [19][20][21]. With the outbreak of COVID-19, accounting for testing has become a critical element in effectively using an SEIR model [22, 23].

3. Model Overview

MechBayes extends the SEIR model framework to account for a joint observation model on incident cases and deaths, time-varying transmission, and time-varying discrepancies in the case fatality ratio due to testing, asymptomatic cases etc. We do this by choosing flexible non-parametric models that capture trends in disease parameters without relying on external data.

4. Data

In this analysis we used confirmed case counts and deaths as reported by the Johns Hopkins University Center for Systems Science and Engineering [24]. This a time series dataset which we truncate to begin March 1st 2020 to August 1st 2020 and captures all

50 states, as well as Guam, Puerto Rico, American Samoa, District of Columbia, Northern Mariana Islands and U.S. Virgin Islands. As noted in [25], COVID-19 cases are under-reported, with the fraction of all infections reported as cases for the U.S. estimated at 20-30% [26].

5. Targets

We made probabilistic forecasts for 1–4 week ahead incident and cumulative deaths for all geographies. An individual forecast distribution is represented by a set of quantiles, $\mathbb{Q} = .01, .05, .10, ..., .90, .95, .99$, with the median (.5 quantile) representing the point-forecast. Since our model produced cumulative forecasts by aggregating incident, we choose to only evaluate incident.

6. MechBayes

In a given time-step (e.g. one day), each member of the population of a single geography belongs to one of the following mutually exhaustive compartments: susceptible S, exposed but not yet infectious E, infectious I, recovered R, hospitalized before death D_1, and deceased D_2 (Figure 2). Here we assume everyone who is hospitalized will eventually become deceased in order to separate the rate of flow into both a case fatality ratio (CFR) parameter as well as a time from symptoms to death parameter, which both have prior estimates from the literature [27]. We omit an explicit hospitalization compartment since the available hospitalization data is highly variable by state and suffers even more reporting issues than case data. For simplicity, we assume a closed population of size N. The following parameters govern how members of the population move between compartments:

- $\beta(t)$: transmission rate, which we allow to vary by time t
- σ: rate of transition from the exposed state E to infectious state I; i.e., $1/\sigma$ is the expected duration of the time between exposure and symptom onset.

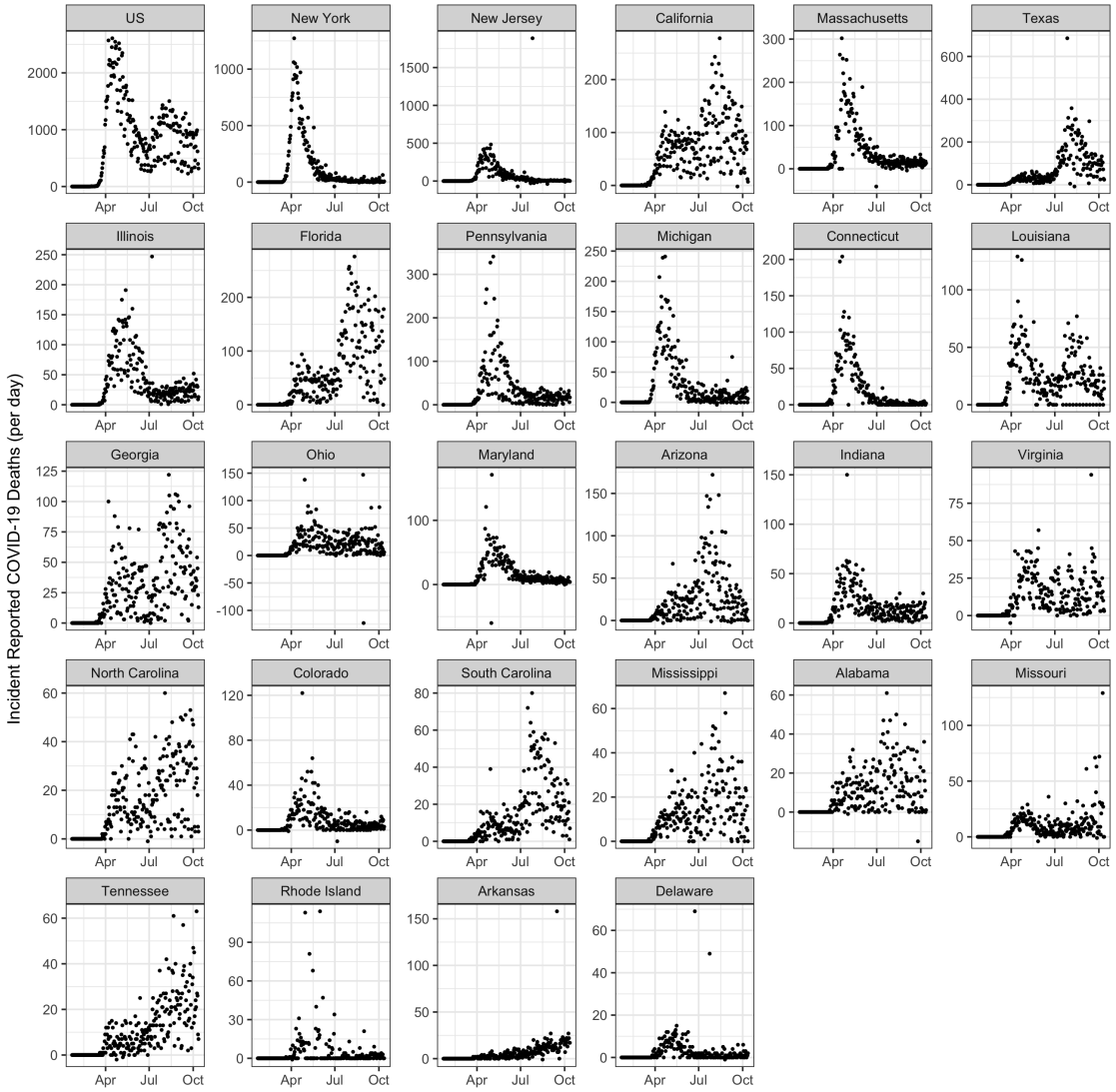


FIGURE 1. Deaths by state for any state with over 50 incident deaths for a given day. Notice the large variability in incident reporting. There are significant delayed reporting, with New Jersey reporting over 1500 back-logged deaths when switching to include "probable" deaths. In some states, there appears to be a weekly cycle where deaths are under-reported on the weekend. This is especially pronounced for the U.S. as a whole. We can also see that there are some negative incident deaths, where data are revised to account for deaths that were incorrectly attributed to COVID-19.

- γ: rate of transition from the infectious state I to no longer being infectious (either to state D_1 or R); i.e., $1/\gamma$ is the expected duration of the infectious period
- ρ: fatality rate (i.e., probability of transitioning from I to D_1 instead of I 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)

For a given time-step t, the following differential equations describe the changes in each compartment:

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta(t)\frac{SI}{N} \\
\frac{dE}{dt} &= \beta(t)\cdot\frac{SI}{N} - \sigma E \\
\frac{dI}{dt} &= \sigma E - \gamma I \\
\frac{dR}{dt} &= (1-\rho)\gamma I \\
\frac{dD_1}{dt} &= \rho\gamma I - \lambda D_1 \\
\frac{dD_2}{dt} &= \lambda D_2 \\
\frac{dC}{d} &= \sigma E
\end{aligned}
\tag{1}
$$

Here, we include the C compartment to be able to observe the cumulative count of new infections. This captures only the flow into I.

We can write this in a state space representation as follows:

$$X(t) = (S(t), E(t), I(t), R(t), D_1(t), D_2(t), C(t))$$

The update from time t to time $t+1$ can be solved numerically as

$$\boldsymbol{X}(t+1) = \mathrm{RK4}\left(\boldsymbol{X}(t), \frac{dX}{dt}, \beta(t)\right) \tag{2}$$

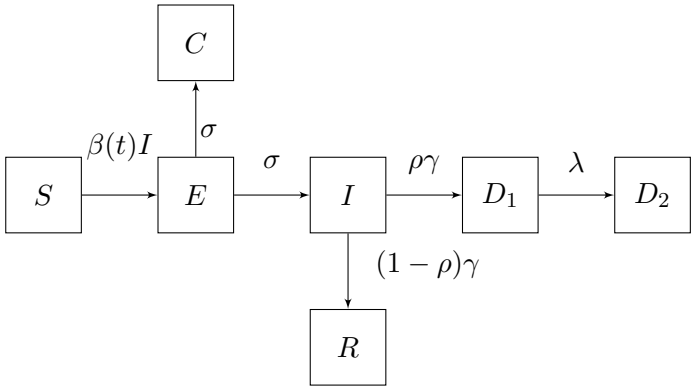


FIGURE 2. Comparmental model parameters

, where RK4 is the Runge-Katta 4th order approximation [28].

6.1. **Time-varying transmission.** We have seen significant efforts to control the spread of COVID through non-pharmaceutical interventions. These include social distancing, lock-downs, and mask wearing. To add to the complexity, these interventions have been implemented and repealed at different time points. They also face compliance issues [29]. In order to capture the aggregate effect of the interventions non-parametrically we choose a flexible model for the time-varying transmission parameter. We allow $\beta(t)$ to vary as follows,

$$log(\beta(t)) \sim N(log(\beta(t-1), \sigma_\beta^2) \tag{3}$$

This model assumes that forecasts are made on the current level of interventions because $\mathbb{E}[log(\beta(t+1))] = log(\beta(t))$. That is, the expected value of a random walk in the forecasting stage is the estimated value at the final time point of available data.

However, this non-parametric model is particularly susceptible to noise in reporting of cases, since $\beta(t)$ is the parameter that takes individuals from S into E. To avoid instability issues, especially when forecasting, we smooth each posterior sample of β to be the average

over the last 10 days from the sample. This is a large enough window to smooth over most reporting issues (excepting delayed reporting).

$$\beta^*_{forecast}(t+k) = \frac{1}{10}\sum_{i=0}^{9}\beta^*(t-i) \tag{4}$$

where $\beta^*(t)$ denotes a sample from the posterior at time t.

6.2. **Observation Model.** The observed data used to fit the model is based on time-series data of incident confirmed cases $Cases_t$ and incident recorded deaths $Deaths_t$. For a given state and day, the change in the confirmed cases and reported deaths are subset of the cumulative number of new infections $C(t)$ and cumulative number of deaths $D2(t)$, respectively. To handle this, we introduce two additional parameters. First, we introduce a case and death deviation, beyond the case fatality ration, of cases $p_{c,t}$ and the case and death deviation model of deaths p_d. For both, we set fairly flat priors to reflect these parameters are poorly determined from observed data.

In more detail, $p_{c,t}$ can be thought of as an aggregate probability of a case being detected and then flowing through the compartments. We assume its prior distribution is given by $p \sim \text{Beta}(15, 35)$, such that $\mathbb{E}[p_c] = 0.3$ with 90% probability between $0.22, 0.38$. This means that we expect 30% of cases to be detected initially, as suggested by the literature [30]. However, we also allow this to vary by time. We do not intend for this to be interpretable as purely reflecting testing, but rather an aggregate measure of testing, reporting issues, and general departure from our prior estimate of the case fatality ratio.

$$logit(p_{c,t}) \sim N(logit(p_{c,t-1}), \sigma^2) \tag{5}$$

We also assume the probability that a COVID-19 death is reported p_d has a prior distribution given by $p_d \sim \text{Beta}(90, 10)$. This prior satisfies $\mathbb{E}[p_d] = 0.9$ with 90% probability

between 0.89,0.92. That is, we assume that deaths due to COVID-19 are most often correctly reported [31].

Using the above SEIR model and these detection probabilities, we can then express the observed incident numbers of confirmed cases and deaths as follows.

$$\text{Cases}_t \sim NB(p_{c,t} * [C_t - C_{t-1}], \sigma_c^2) \tag{6}$$

$$\text{Deaths}_t \sim NB(p_d * [D_{2_t} - D_{2_{t-1}}], \sigma_d^2) \tag{7}$$

Where the difference in $C_t - C_{t-1}$ allows us to translate cumulative new cases to incident cases and similarly with deaths.

6.3. **Epidemiological Model Parameters.** We use relatively informative priors for epidemiological parameters, such as γ, σ, ρ, λ, and initial compartment values. The details are described in the Appendix A1. However, the identifiability of model parameters in compartmental models where the data consists only of a time series of incident cases and deaths presents a problem for uninformative priors. Using the renewal style equations, it can be shown that the number of newly infected at time t is a function of the time-varying reproductive number, serial interval and previously reported new infections [32]. This means that a single time series does not contain enough information to separately estimate both the serial interval and the time-varying reproduction number. In an SEIR model, the serial interval is distributed exponential with rate parameter $\sigma + \gamma$ [32]. Additionally, the time varying reproduction number is $R_t = \frac{\beta(t) * S(t)}{\gamma}$. Therefore, the time series of incident cases is not enough to uniquely identify $\gamma, \sigma, \beta(t)$. In order to make the model identifiable, we impose tight priors on the parameters σ and γ as estimated by the literature (in essence fixing the serial interval), and we let $\beta(t)$ vary freely. This reflects the underlying biology of the system, since the reciprocal of the sum of σ and γ may be interpreted as the average

time from when an individual becomes infected to when they infect someone else, given that they infect someone else. This is a biological property of the disease, rather than $\beta(t)$ which contains both the biological transmissibility as well as the aggregate effects of human behavior through intervention. This highlights a fundamental philosophical difference between using compartmental models for forecasting rather than interpreting parameters for epidemiological purposes. However, putting relatively informative priors on σ and γ, instead of fixing them, still allows for variation by state due to differing demographic characteristics such as age structure. Fitting the model in a Bayesian way allows for this unique trade off.

6.4. **Fitting.** We use the Hamiltonian Monte Carlo algorithm implemented in numpyro to fit the model to data [33]. That is, given a time series of confirmed cases ($\text{Cases}_{1:t}$) and confirmed deaths ($\text{Deaths}_{1:t}$) we use Bayesian inference (via HMC) to obtain

$$f(\boldsymbol{\theta}|\text{Cases}_{1:t}, D_{1:t}) \propto f(\text{Cases}_{1:t}, \text{Deaths}_{1:t}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{8}$$

where $\boldsymbol{\theta}$ is a vector containing all model parameters.

$$\boldsymbol{\theta} = [\beta_t, \sigma, \gamma, \rho, \lambda, p_{c,t}, p_d, \sigma_c^2, \sigma_d^2, I_0, E_0, D_{1_0}, D_{2_0}, R_0] \tag{9}$$

We draw 1000 warm-up sample and then 1000 posterior samples of model parameters. This allows us to forecast using the posterior predictive distribution.

$$f(C_{t:(t+h)}, D_{t:(t+h)}) = \int_\theta f(C_{1:t}, D_{1:t}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{10}$$

For dan to fill in

7. Experimental Setup

To evaluate our model, we examine two different scenarios. First, we describe the submission process and infrastructure used for the real time evaluation as part of the U.S. Centers for Disease Control (CDC) COVID-19 Forecast Hub consortium. Second, we describe the internal evaluation used to demonstrate our model enhancements improve accuracy over a naive compartmental model.

7.1. **CDC Forecast Evaluation.** We began submitting forecasts to the CDC for incident deaths on May 10th 2020 and have since submitted forecasts every Monday from then until August 1st 2020. The forecasts use daily data up to and including the Sunday before submission the next Monday. The one week ahead forecast corresponds to the following Saturday, the two week ahead to the second following Saturday and so on. We use the model submissions made in real-time evaluated on both MAE and coverage probability for 13 submissions. We subset to the 50 states and Washington D.C., which is the largest set of locations where forecasts were made for each date.

In the real-time evaluation we also made manual adjustments to account for delayed reporting through a quality-assurance process. This involved,

- Identifying outliers in recently reported incident cases.
- Searching for documented evidence of a data dump. These are usually recorded on state department of health websites and sometimes local news outlets.
- Manually redistributing the incident deaths evenly over the time-frame mentioned by the department of health or news outlet for the backlog window.

This process ensured that the observed data does not contain any identifiable outliers (meaning documented by outside sources). In real-time this is necessary to avoid drastic over-predictions caused by delayed reporting. For example, New Jersey reported nearly 1,600 daily deaths as it switched from reporting only confirmed deaths from COVID-19, to confirmed and probable on 2020-06-25. This would have caused a drastic increase in

predictions if not properly identified as data dump. On 2020-07-07 Texas removed 3,000 confirmed cases when they discovered the reported cases were a result of anitgen testing, which were not considered reportable.

We compare the results against the CDC baseline model. This is a purely statistical model that uses historical COVID-19 deaths only. At all forecast horizons, the median of the forecast distribution from the baseline model is equal to the most recent reported incidence. The model obtains a non-parametric distribution around this median by adding and subtracting past observed differences in incidence from one week to the next for a specific location. This model is fit state by state.

7.2. **Ablation Test.** While real-time model evaluation is valuable for understanding evolving model performance, we also perform a retrospective evaluation using three model variants to demonstrate the improvement in accuracy over a baseline SEIR model. We define the following variations on MechBayes,

- **MechBayes Case/Death Time-Varying** Mech Bayes as submitted to the CDC. That is, a model using negative binomial observation noise as well as a time-varying random walk, using a joint likelihood over cases and deaths.
- **MechBayes Case/Death Fixed** MechBayes Case/Death Time-Varying with $p_{c,t}$ fixed to p_c, that is, removing the time-varying case and death deviation model.
- **MechBayes Death Fixed** MechBayes Case/Death Time-Varying with observations on cases removed.

Note that these are nested models, with MechBayes Death Fixed contained in MechBayes Case/Death Fixed contained in MechBayes Case/Death Time-Varying.

We also fix all non-model component variation. That is, we average over the last 10 days of $\beta(t)$ when forecasting, as well as manually redistributing delayed reporting. This ensures that the comparison is only on model components, and not on data discrepancies. Note that we do not include a model without a time varying transmissibility parameter.

This is because such a model would assume no interventions were put in place, which clearly violates the data-generating process. Previous Covid-19 modeling attempts have established that time-varying transmissibility is essential [23] [34][21] [16].

8. Results

8.1. **Model Fitting and Inference.** MechBayes is able to accurately model the observed data, adapting to highly variable incident death reporting, variable transmission rates, and overall heterogeneity of incidence curves (Figure 3 A). The model is able to adapt to highly variable incident death reporting, variable transmission rates, and overall heterogeneity of incidence curves. The model is also able capture the uncertainty of the differential equation parameters well enough to produce well calibrated prediction intervals. Figure 3 shows prediction intervals at the 95% level, with 92.3% of observations falling within the bounds for each state. However, we can also see that in some states, such as California and Florida, the model is biased high, with almost all observations outside of the 95% prediction interval falling below. Figure 3 also shows 4 weeks of daily forecasts, along with the daily observed incidence for 1 week out. We can see that the predictions are tracking the data even under the highly variable weekly reporting fluctuations.

MechBayes is able to adapt to changes in transmission through the $\beta(t)$ parameter (Figure 3 C). Transmission was high across all four example states in March, with exponential growth rates shown in panel B. Transmission then slowed significantly in April across each states. While New York remained low through July, Texas, California and Florida saw increases in late June. This is reflected in the increase in incident deaths approximately two weeks later in mid-July.

MechBayes is able to adapt to changes in the case-fatality ratio through the $p_{c,t}$ parameter (Figure 3 D). At the start of the pandemic, only those who were very sick were able to be tested. This lead to an unusually high case-fatality ratio (with respect to estimates from other countries). MechBayes was able to increase the number of latent cases (by a

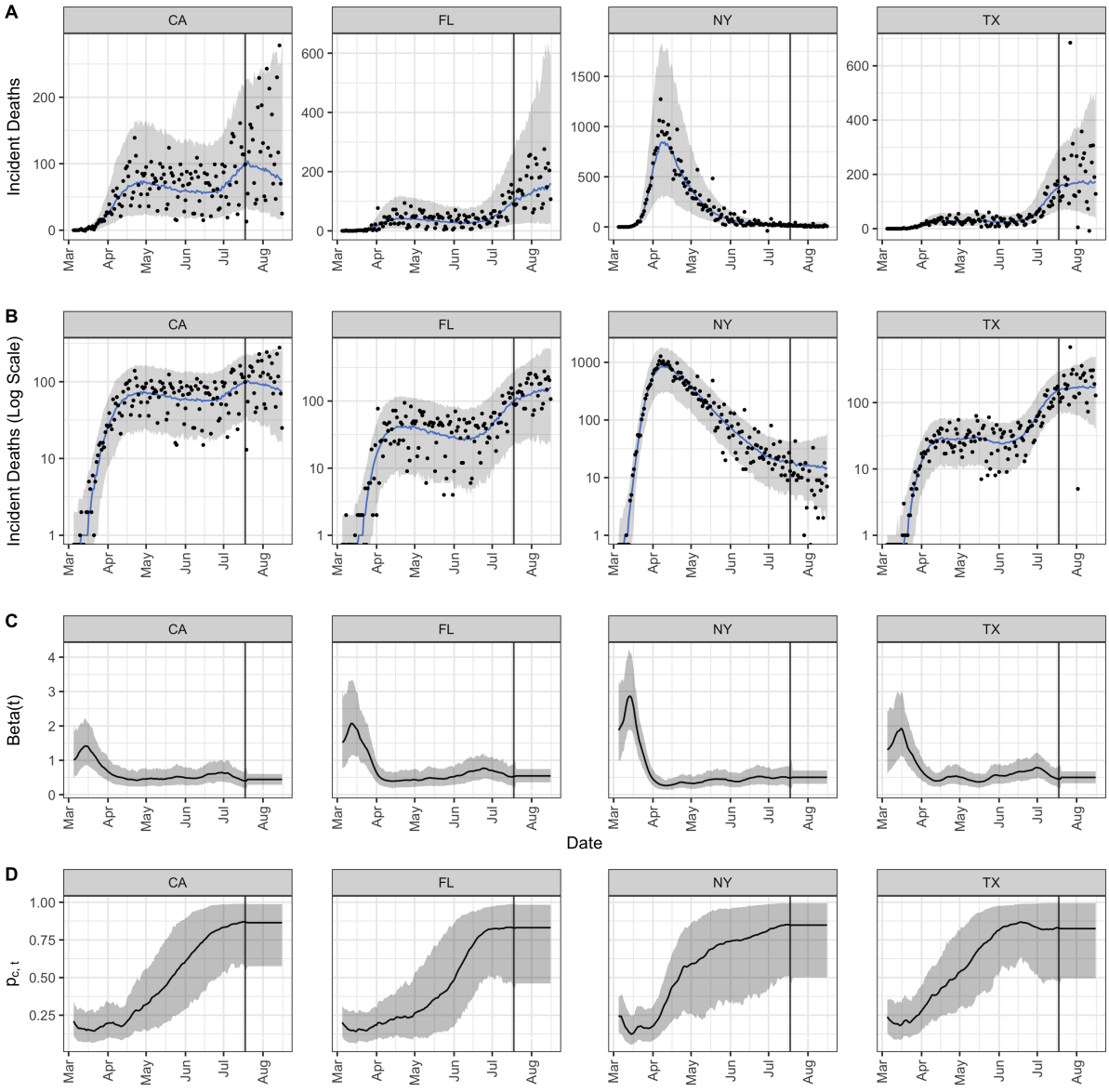


FIGURE 3. **A,B** Example fit and forecast for 2020-07-19 for four selected states. Grey bands represent 95% prediction intervals. Blue line represents median forecast. MechBayes is able to produce well calibrated fits to the data as well as accurately tracking trends in incident deaths. **B** Posterior of β_t for each of the four states, with 95% credible intervals in blue. Note that this parameter does not reflect interventions specifically, but rather any change in transmissibility. **C** Posterior of the case-death discrepancy for the four example states. Here we can see that over time, the number of observed cases that resulted in deaths increased. This does not necessarily reflect testing, but an overall increase in the ability to observe cases as compared to the predicted number of cases one would obtain through the case-fatality ratio.

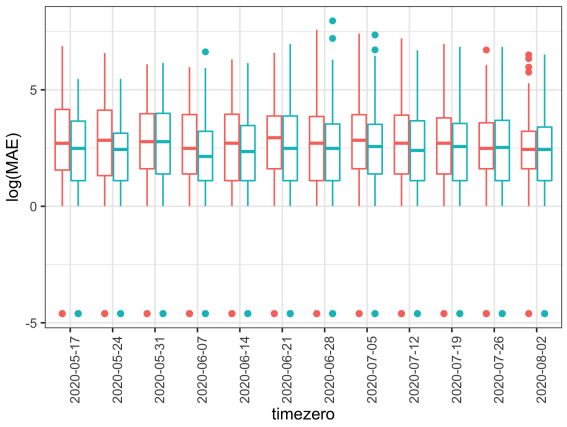
factor of $p_{c,t}$) to lower the estimated case-fatality ratio and avoid extreme over estimates of incident death predictions that would have accompanied such a drastic increase in cases in March. The model was able to adapt to this change across all four example states.

8.2. **CDC Forecast Results.** We next turn to the comparison of MechBayes against the CDC baseline model. This baseline model uses the previous daily incident as the mean forecast for the current daily incidence, along with bootstrapped prediction intervals from historical changes in daily incidence. See CDC COVID-19 Forecast Hub for more details [1].

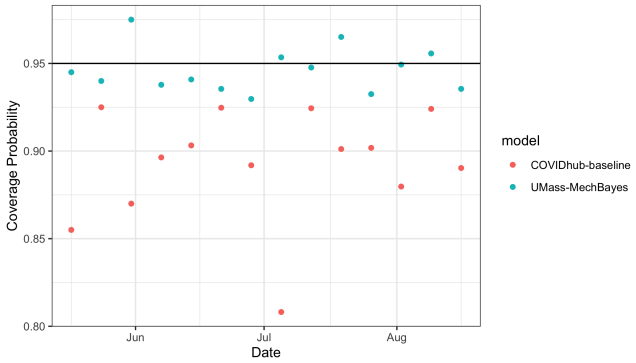
Overall, MechBayes had an MAE of 34.1, when averaging over all regions, forecast dates, and targets. The CDC baseline model had an MAE of 53.2. Similarly, the empirical coverage probability at the 95% level for MechBayes was 94.6%, compared to 93.2% for the baseline model.

We also break down the results by week the forecast was made (forecast week) and averaging over both region and target. As we can see from Figure 4 (A,B), MechBayes outperformed the CDC baseline on MAE and coverage probability when broken down by forecast week. While the improvements broken down by time are small, there is an improvment in 9 out of 12 evaluation weeks in MAE. As we can see from 1, in weeks with the largest increase in deaths (mostly during the month of May) MechBayes significantly outperformed the baseline model. However, in weeks with a small increase or a decrease in deaths, the scores were much closer. This suggests that MechBayes performs well where it counts, when the epidemic is taking off nearly exponentially.

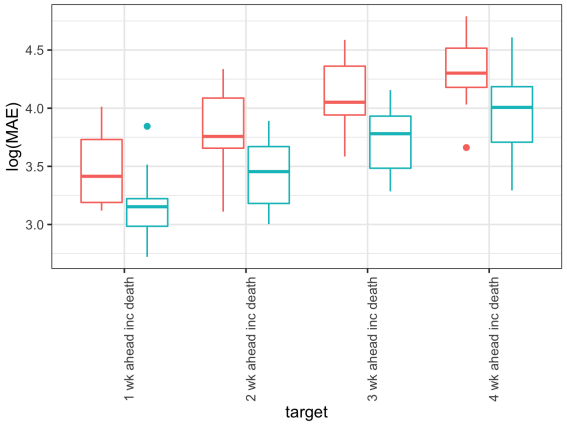
We also break down the results by geographical region, as seen in Figure 4 (C,D). Here we see consistent improvements in MAE under MechBayes for regions with high total death counts with an improvement in the average MAE in 16 out of the 20 states with the highest total death count. Out of the states with the 10 highest death counts, only in California did the baseline model outperform MechBayes on the average MAE. We break



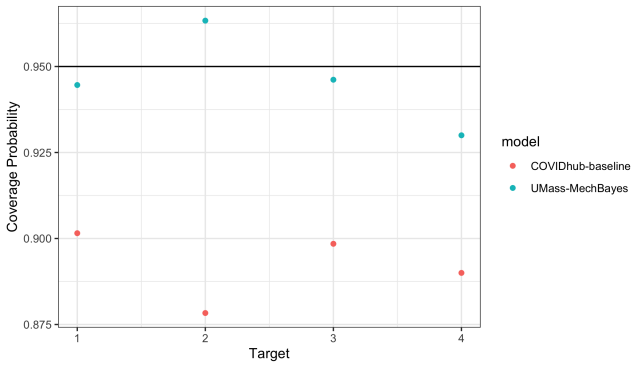
(A) MAE by forecast week



(B) Coverage probability by forecast week

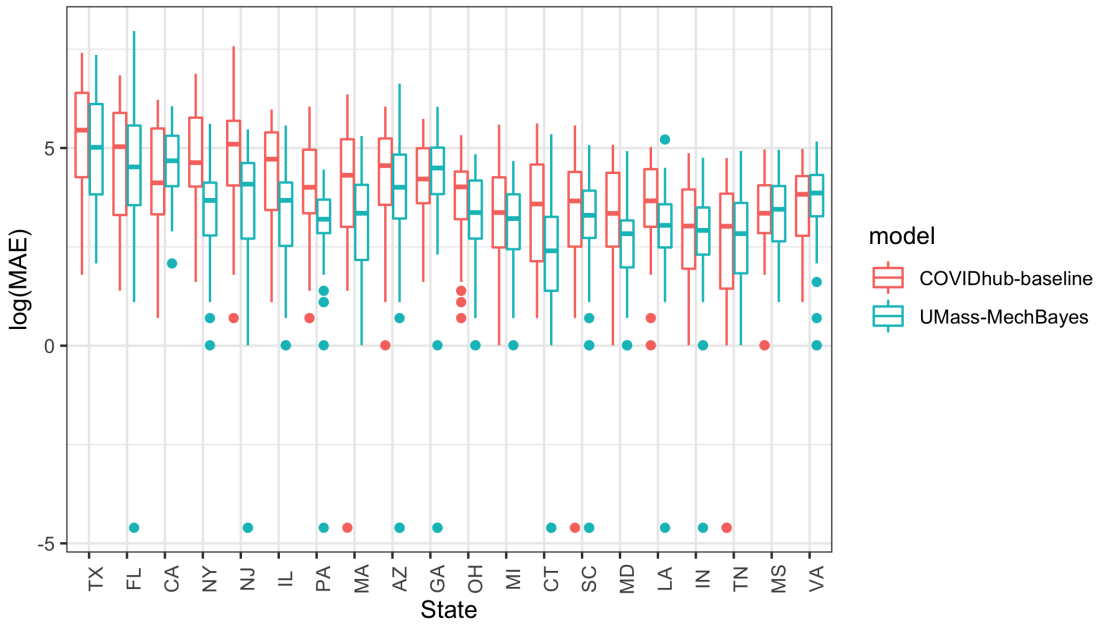


(C) MAE by target

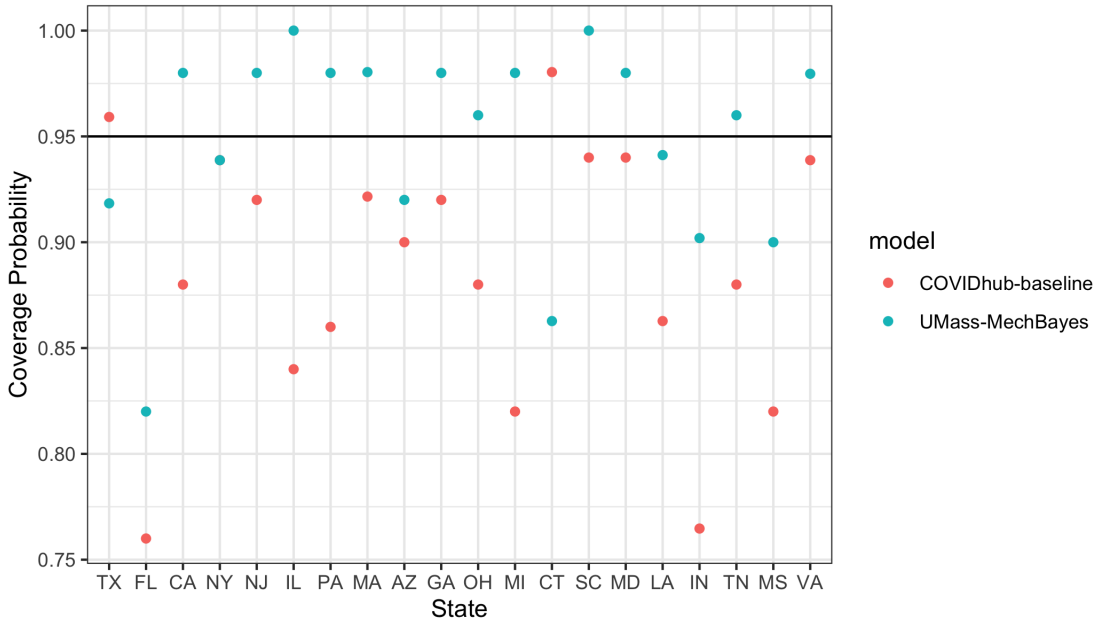


(D) Coverage probability by target

FIGURE 4. Scores from CDC COVID-19 forecast initiative broken down by target and forecast week. In Panels A and B, each observation represents mean absolute error (MAE) and empirical prediction interval coverage rate averaged across all targets and locations for the given week in which forecasts were made. MechBayes improves in average MAE in 9 out of 12 timepoints. MechBayes improves in MAE across all targets. We can see that as the horizon increases from 1 to 4 weeks ahead, MAE, reflecting an increase in difficulty of forecasting further ahead in time. We also see a uniform improvement in coverage probability when broken down by timezero and target. Here the horizontal line represents 95% coverage interval.



(A) MAE by region



(B) Coverage probability by region

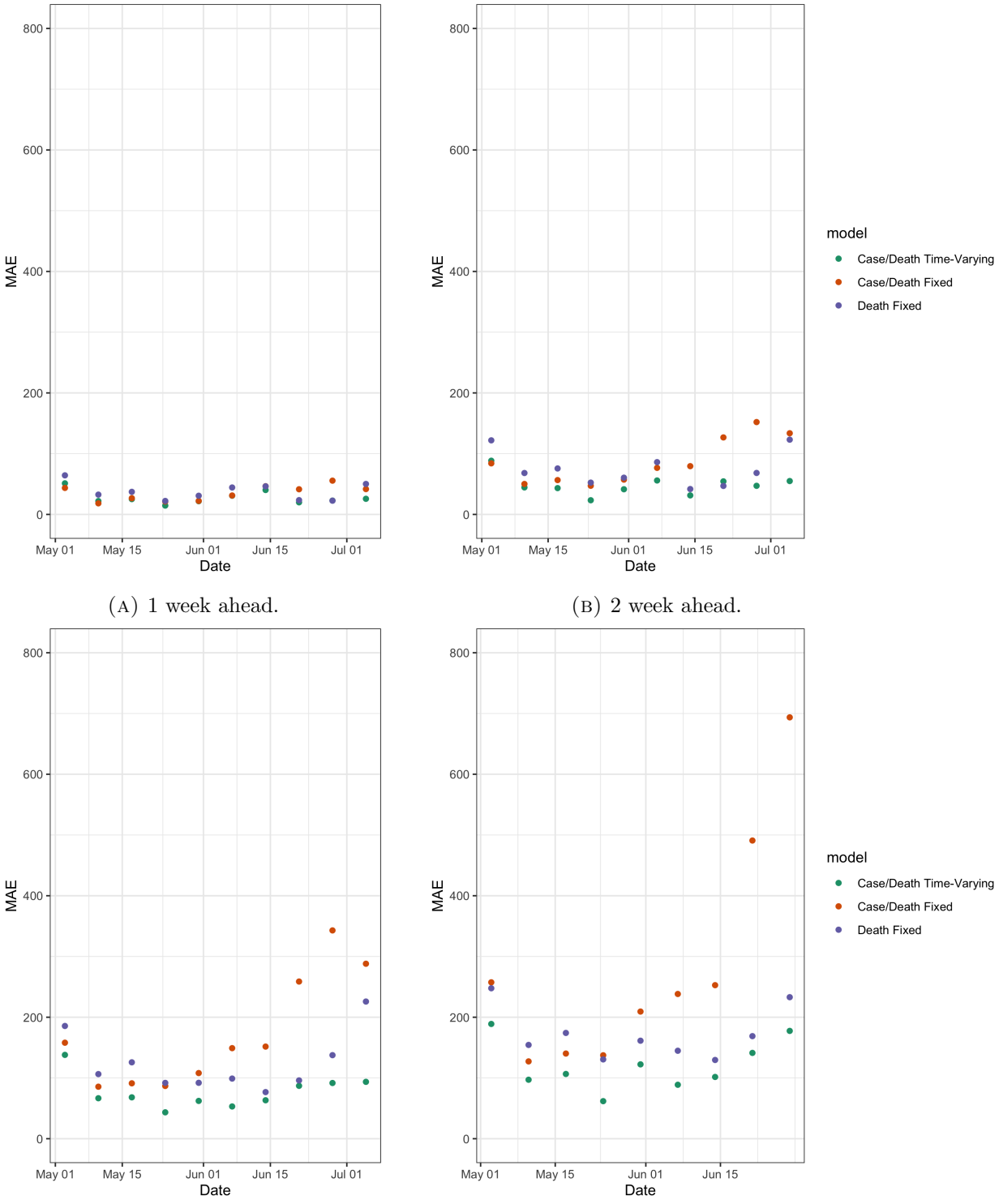
FIGURE 5. Scores from CDC COVID-19 forecast initiative broken down by region for the 20 regions with the highest incident deaths. Each observation represents mean absolute error (MAE) or empirical prediction interval coverage rate averaged across all targets and forecast dates for the given region in which forecasts were made. Regions are sorted by total deaths left to right. Here we can see that MechBayes has the greatest increase in states with large total death counts. States with medium or small death counts have more mixed results. The improvements in coverage probability are more mixed, with MechBayes improving coverage probability in 12 of 20 regions.

down the results by target by averaging over region and forecast week 4 (E,F) Here we can see uniform improvement over the baseline model by MechBayes in terms of MAE and coverage probability. We can also see that the MAE increase as horizon increase, which is to be expected. We can also see that incident MAE is lower than cumulative MAE, which is again to be expected due to the lower absolute numbers of incident deaths.

Finally, we include a formal test of the difference in MAE to demonstrate statistically significant advantages in using MechBayes over the baseline model. MechBayes has statistically significantly lower MAE for 1-4 weeks ahead at the 95% level (coefficients in Appendix 2). The difference in skill seems to improve as the forecast horizon increases. For example, if the baseline model MAE for a particular 1 step ahead forecast is 30, one would expect the MechBayes forecast to have an MAE of 24.8. If the baseline model MAE for a particular 4 step ahead forecast is 30, one would expect the MechBayes forecast to have an MAE of 22.5. We demonstrate that the mixed model proposed above fits the MAE results reasonably in Appendix 2.

8.3. **Ablation Test Results.** MechBayes Case/Death Time-Varying is consistently better than MechBayes Death Fixed or MechBayes Case/Death Fixed. The only exception seems to be June 22nd 2020 for the two week ahead target. The difference in MAE between MechBayes Case/Death Time-Varying and the competing models increases as forecast horizon increases. This suggest that MechBayes Case/Death Time-Varying is a better long term forecasting model.

We can also see that the MechBayes Death Fixed is consistently better than MechBayes Case/Death Fixed. This may be evidence that naively including case data, without adjusting for discrepancies between cases and deaths due to testing etc., may be worse than not including it at all.



(A) 1 week ahead.

(B) 2 week ahead.

(C) 3 week ahead.

(D) 4 week ahead.

FIGURE 6. Ablation test MAE broken down by forecast week and target. We can see that MechBayes Case/Death Time-Varying performs better than MechBayes Case/Death Fxied and MechBayes Death Fixed in almost all breakdowns. We can also see that the improvement becomes more pronounced at larger horizons, suggesting that MechBayes Case/Death Time-Varying is a better longer term forecasting model.

9. Discussion

MechBayes is a fast, fully Bayesian compartmental model capable of accounting for real-world data challenges during a pandemic. This model produced consistently accurate real-time forecasts over the course of 3 months, and was ranked as one of the top 3 of 11 models on an independent scoreboard of COVID-19 forecast models [35]. 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**. As Figure 4 shows, MechBayes improved relative to the baseline model when broken down by timezero and target. The results are more pronounced when breaking down by target, where the boxplots are almost completely separated. These results are statistically significant for 1-4 weeks ahead, with larger improvements at later horizons, suggesting that the biggest improvements from MechByes over the baseline model comes in longer term forecasting. Additionally, as seen in Figure 5 , the biggest gains in performance in MAE occur in regions with the largest total deaths counts. This is a desirable feature of a pandemic forecasting model.
- **MechBayes is probabilistically well-calibrated**. As we can see from Figure 4 (B,D), the coverage probability of MechBayes is much closer to 95%. When averaging over target, region, and timezero the coverage probability is 94.6%. This suggests that MechBayes is very well calibrated at the 95% level. The two sources of uncertainty in MechBayes come from the distribution over the differential equation parameters and the observation noise. The model is able to learn reasonable estimates of these variance parameters to produce well calibrated results.
- **Adding case data when predicting deaths is helpful but only when accounting for data quality issues**. Our ablation tests (Figure 6) clearly shows that time-varying case and death deviation model is a key feature in the model

for reducing MAE of forecasts. The "Full" model that both incorporates incident cases into the model likelihood and allows for a flexible deviation between cases and deaths is consistently more accurate than a model that does not account for cases at all ("Death Only" model) and a model that does account for cases but does not account for a time-varying deviation between cases and deaths.

- **Most epidemiological parameters are unidentifiable from the data**. MechBayes requires relatively informative priors on σ, γ, ρ and λ. These parameters reflect the biology of the disease, from latent incubation period to average time from symptom onset to death. Since the data used to fit MechBayes is only a time series of confirmed cases and deaths, these parameters are simply not identifiable from the data. MechBayes was built for forecasting, not inference on epidemiological parameters. However, using a Bayesian framework we can simultaneously set priors based on the literature and allow for small deviations from the prior due to variations across states.
- **Allowing for time-varying transmissibility is necessary to non-parametrically capture the effect of interventions**. Our ablation test explicitly did not include a model that fixed β across time. This is because the model would not converge without the flexibility to capture changes in transmission. While non-parametrically modeling interventions is appealing from a forecasting perspective, it does modify the philosophy behind compartmental modeling. By including such a flexible parameter, we may view MechBayes as simply a random-walk model, with a set of epidemiological parameters transforming that random-walk in an almost deterministic way to match both cases and deaths. For instance, if the variance of the random walk σ_β^2 was allowed to be arbitrarily large, then $\beta(t)$ could vary enough to match the data exactly. This would clearly attribute reporting issues as

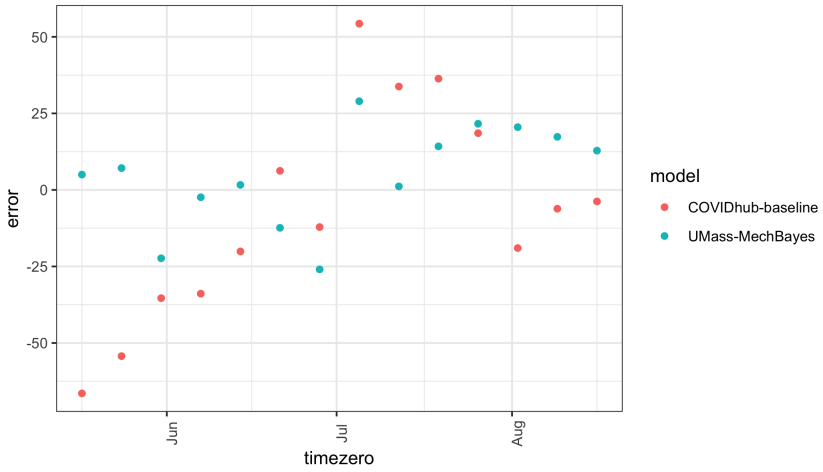


FIGURE 7. Bias of MechBayes and COVID-Baseline as a function of time.

true changes in transmissibility. Bypassing the epidemiological interpretation that compartmental models provide.

10. CONCLUSION

We have seen that MechBayes is a powerful Bayesian compartmental model that can capture the real-world complexities of forecasting during a pandemic. Through real-time and retrospective evaluation, we have demonstrated the success of MechBayes in forecasting COVID-19. The model is able to improve over a naive baseline model as well as a naive compartmental model. Allowing for time-varying interventions and case and death deviation model is a necessary model component during a real-time pandemic forecasting effort.

While we chose an exponential random walk on $\beta(t)$ there are many other choices for flexible non-parametric modeling of transmissibility. Further work might consider a spline model, or a Gaussian process, or semi-parametric models capable of taking intervention dates as covariates.

However, using relatively simple non-parametric methods we were able to beat a baseline model in both point and probabilistic forecast evaluations. Our ablation tests show that extending the basic SEIR compartmental to real-world pandemic challenges improves forecasting accuracy.

11. Appendix

11.1. **Statistical Test for Significance.** In order to do this, we just a random effects regression model of the form,

$$\begin{aligned} log(MAE_{m,t,r,h}+1) &= \beta_0 + \beta 1 * h_1 + ... + \beta_3 * h_3 \\ &+ \beta_4 * h_1 * mb + ...\beta_8 * h_4 * mb \\ &+ b_r + \epsilon \\ b_r &\sim N(0, \Sigma_b^2) \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

where mb is an indicator for the MechBayes model, t is timezero, r is region, and h is target horizon (1-4 week ahead). We chose this model because it explains the variation in MAE by model and horizon while allowing varying baseline MAE values by region. Here, variation over time in MAE within a specific region is explained by differences in model performance This leads to the following coefficient estimates for the fixed effects.

	Estimate	Std. Error	df	t value	Pr(>|t|)
β_0	2.47	0.17	54.87	14.97	0.00
β_{h_1}	0.30	0.06	4643.00	5.18	0.00
β_{h_2}	0.44	0.06	4643.00	7.56	0.00
β_{h_3}	0.59	0.06	4643.00	9.79	0.00
$\beta_{h_1 mb}$	-0.19	0.06	4643.00	-3.19	0.00
$\beta_{h_2 mb}$	-0.29	0.06	4643.00	-4.87	0.00
$\beta_{h_3 mb}$	-0.23	0.06	4643.00	-3.94	0.00
$\beta_{h_4 mb}$	-0.29	0.06	4643.00	-4.72	0.00

Table 1. Coefficient estimates and t-values for MAE evaluation model. We can see that MechBayes performs statistically significantly better than the baseline model for 1-4 weeks ahead. The performance increase seems to grow as horizon increases.

11.2. **Seeding Epidemic.** Due to the under-reporting of cases, we cannot use the observed data to seed the epidemic. We instead allow the model to find the initial state values for all compartments except the number of susceptible people, which we take as the population size of the geographic region minus the sum of the initial values for the other compartments to enforce the constraint that the entire system size sums to the population size. We do this by assigning uniform probability to all initial states where the number of people in any given compartment at time zero does not exceed 2% of the total population. This is a highly conservative estimate for the number of infected, exposed, dead and recovered people at the start of the epidemic which is most likely much lower than 2% of the population.

$$
\begin{aligned}
E_0 &\sim \text{Unif}(0, 0.02N) \\
I_0 &\sim \text{Unif}(0, 0.02N) \\
D_{1_0} &\sim \text{Unif}(0, 0.02N) \\
D_{2_0} &\sim \text{Unif}(0, 0.02N) \\
R_0 &\sim \text{Unif}(0, 0.02N)
\end{aligned}
$$

This allows us to initialize the process model:

(11)

$$X(0) = (S(0), E(0), I(0), R(0), D_1(0), D_2(0), C(0)) = (N - E_0 - I_0 - D_{1_0} - D_{2_0}, E_0, I_0, R_0, D_{1_0}, D_{2_0}, I_0)$$

11.3. **Priors.** We also place the following priors on the transition parameters:

$$\sigma \sim \Gamma(5, 5\hat{d}_E)$$

$$\gamma \sim \Gamma(7, 7\hat{d}_I)$$

$$\beta(0) \sim \Gamma(1, \hat{d}_I/\hat{R})$$

$$\rho \sim \text{Beta}(10, 90)$$

$$\lambda \sim \Gamma(10, 100)$$

Our prior on rate for leaving the exposed compartment σ satisfies $\mathbb{E}[\sigma] = 1/\hat{d}_E$, where $\hat{d}_E$ is an initial guess of the duration of the latent period. Currently, we assume $\hat{d}_E = 4.0$ based on published estimates (shortened slightly to account for possible infectiousness prior to developing symptoms) [30]. Our prior on the rate for leaving the infectious compartment γ satisfies $\mathbb{E}[\gamma] = 1/\hat{d}_I$, where $\hat{d}_I$ is an initial guess for the duration of infectiousness. The current setting is $\hat{d}_I = 2.0$ to model the likely isolation of individuals after symptom onset [36]. Our prior on the initial transmission rate is derived from the relationship between the basic reproductive number $R(0)$ and the length of the infectious period: $R(0) = \beta(0)/\gamma = \beta(0) \times \hat{d}_I$. Therefore, we set our prior on the initial transmission rate to satisfy $\mathbb{E}[\beta(0)] = \hat{R}/\hat{d}_I$ where $\hat{R} = 3.0$ is an initial guess for $R(0)$ and $\hat{d}_I = 2.0$, as described above. Our prior on the fatality rate ρ satisfies $\mathbb{E}[\rho] = 0.1$ with 90% probability of being between

$$0.06, .14$$

. Finally, our prior on the rate at which dying patients succumb satisfies λ $\mathbb{E}[\lambda] = 0.1$ with shape 10 corresponding to roughly 10 days in the D_1 compartment.

References

[1] ReichLab. *COVID-HUB*, 2020 (accessed July 27, 2020). https://doi.org/10.5281/zenodo.3963372.

[2] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *medRxiv*, 2020.

[3] Chelsea S Lutz, Mimi P Huynh, Monica Schroeder, Sophia Anyatonwu, F Scott Dahlgren, Gregory Danyluk, Danielle Fernandez, Sharon K Greene, Nodar Kipshidze, Leann Liu, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health*, 19(1):1659, 2019.

[4] Monica F Myers, DJ Rogers, J Cox, Antoine Flahault, and Simon I Hay. Forecasting disease risk for increased epidemic preparedness in public health. In *Advances in parasitology*, volume 47, pages 309–330. Elsevier, 2000.

[5] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[6] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[7] Dave Osthus, Kyle S Hickmann, Petruţa C Caragea, Dave Higdon, and Sara Y Del Valle. Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics*, 11(1):202, 2017.

[8] Jimmy Boon Som Ong, I Mark, Cheng Chen, Alex R Cook, Huey Chyi Lee, Vernon J Lee, Raymond Tzer Pin Lin, Paul Ananth Tambyah, and Lee Gan Goh. 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(4):e10036, 2010.

[9] IM Hall, R Gani, HE Hughes, and S Leach. Real-time epidemic forecasting for pandemic influenza. *Epidemiology & Infection*, 135(3):372–385, 2007.

[10] Phenyo E Lekone and Bärbel F Finkenstädt. Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006.

[11] Benjamin Bokler. Chaos and complexity in measles models: a comparative numerical study. *Mathematical Medicine and Biology: A Journal of the IMA*, 10(2):83–95, 1993.

[12] Side Syafruddin and MSM Noorani. Seir model for transmission of dengue fever in selangor malaysia. *IJMPS*, 9:380–389, 2012.

[13] Luiz K Hotta. Bayesian melding estimation of a stochastic seir model. *Mathematical Population Studies*, 17(2):101–111, 2010.

[14] Vanja Dukic, Hedibert F Lopes, and Nicholas G Polson. Tracking epidemics with google flu trends data and a state-space seir model. *Journal of the American Statistical Association*, 107(500):1410–1426, 2012.

[15] Gianluca Frasso and Philippe Lambert. Bayesian inference in an extended seir model with nonparametric disease transmission rate: an application to the ebola epidemic in sierra leone. *Biostatistics*, 17(4):779–792, 2016.

[16] Alexandra Smirnova, Linda deCamp, and Gerardo Chowell. Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the seir model. *Bulletin of mathematical biology*, 81(11):4343–4365, 2019.

[17] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, pages 1–6, 2020.

[18] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. *Journal of Thoracic Disease*, 12(3):165, 2020.

[19] Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *arXiv preprint arXiv:2004.04741*, 2020.

[20] Kiesha Prem, Yang Liu, Timothy W Russell, Adam J Kucharski, Rosalind M Eggo, Nicholas Davies, Stefan Flasche, Samuel Clifford, Carl AB Pearson, James D Munday, et al. The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in wuhan, china: a modelling study. *The Lancet Public Health*, 2020.

[21] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, pages 1–5, 2020.

[22] Leonardo López and Xavier Rodo. 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*, 2020.

[23] Sen Pei, Sasikiran Kandula, and Jeffrey Shaman. Differential effects of intervention timing on covid-19 spread in the united states. *medRxiv*, 2020.

[24] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020.

[25] Steven G Krantz and Arni SR Srinivasa Rao. 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. *Infection Control & Hospital Epidemiology*, pages 1–8, 2020.

[26] T Russel, Joel Hellewell, S Abbot, et al. Using a delay-adjusted case fatality ratio to estimate under-reporting. *Available at the Centre for Mathematical Modelling of Infectious Diseases Repository, here*, 2020.

[27] Timothy W Russell, Joel Hellewell, Christopher I Jarvis, Kevin Van Zandvoort, Sam Abbott, Ruwan Ratnayake, Stefan Flasche, Rosalind M Eggo, W John Edmunds, Adam J Kucharski, et al. 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. *Eurosurveillance*, 25(12):2000256, 2020.

[28] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

[29] Andrey Simonov, Szymon K Sacher, Jean-Pierre H Dubé, and Shirsho Biswas. The persuasive effect of fox news: non-compliance with social distancing during the covid-19 pandemic. Technical report, National Bureau of Economic Research, 2020.

[30] Midas. Covid19 parameter estimates. `https://github.com/midas-network/COVID-19/tree/master/parameter_estimates/2019_novel_coronavirus`, 2020.

[31] Daniel M Weinberger, Jenny Chen, Ted Cohen, Forrest W Crawford, Farzad Mostashari, Don Olson, Virginia E Pitzer, Nicholas G Reich, Marcus Russi, Lone Simonsen, et al. Estimation of excess deaths associated with the covid-19 pandemic in the united states, march to may 2020. *JAMA Internal Medicine*, 2020.

[32] Jacco Wallinga and Marc Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007.

[33] Uber AI Labs. *numpyro*, 2020 (accessed July 27, 2020). `https://readthedocs.org/projects/numpyro/downloads/pdf/stable/`.

[34] Sam Abbott, Joel Hellewell, Robin N Thompson, Katharine Sherratt, Hamish P Gibbs, Nikos I Bosse, James D Munday, Sophie Meakin, Emma L Doughty, June Young Chun, et al. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

[35] Youngang Gu. Covid-19 projections. `https://covid19-projections.com/about/#historical-performance`, 2020.

[36] Joseph Heffner, Marc Lluis Vives, and Oriel FeldmanHall. Emotional responses to prosocial messages increase willingness to self-isolate during the covid-19 pandemic. 2020.