MECHANISTIC BAYESIAN FORECASTS OF COVID19

GRAHAM C. GIBSON, NICHOLAS REICH, DAN 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, data quality during a pandemic suffers from under-reporting, delayed reporting of cases and deaths due to reporting infrastructure issues, and limited testing. 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 in forecasting in real time. This includes non-parametric modeling of non-pharamacuetical interventions, time-varying rates of testing for the disease, and joint observations on new case counts and new deaths. The model has been used to submit forecasts to the US Centers for Disease Control, through the COVID19-ForecsatHub repository organized by the Reich Lab. We examine the performance relative to a baseline model in addition to performing a 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 developed into the largest pandemic the country has seen in over a century. Understanding the future trajectory of the pandemic is crucial for minimizing the impact across the nation in terms of healthcare

burden, economic impact, treatments and public health response. Forecasts of incident and cumulative deaths due to COVID may help in resource allocation, vaccine clinical trial planning, and re-opening strategies. 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 [**?**][**?**]. 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 differential equation models to explain the underlying transmission of a disease through a population. First introduced by Kermack and McKendrick, the model assumes the each individual is in one of a mutually exclusive set of compartments, typically either the susceptible, exposed, infected, or recovered compartment [**?**]. The 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 properties in real-time at both local and global scales

.

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 [**?**][**?**][**?**], even including a retrospective model evaluation of the 1918 influenza pandemic [**?**]. Compartmental models have been used not just in respiratory disease but in Ebola [**?**], measles [**?**], dengue [**?**] and a wide variety of other diseases

In this work, 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 data dumps, and both the addition and removal of control measures such as social distancing, lockdown, and mask use. The case and death data is reported as new cases, instead of prevalent, and new deaths, instead of cumulative, which we handle using an observation model on top of the underlying SEIR dynamics. Finally, we choose a Bayesian framework that allows for uncertainty in the epidemiological parameters that are unidentifiable from the data, introducing flexibility suited to forecasting. This also allows for implementation in a cutting edge probabilistic programming framework for speed and accuracy.

. Our main goal in this work 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. However, our m We are not interested in identifying internal parameters of the model, many of which are poorly determined or not identifiable from the available data. We are willing to accept biologically unrealistic parameter values or priors for latent variables; but not ones that lead to gross pathologies during inference and forecasting.

MechBayes is able to account for these operational concerns when forecasting in real-time. We demonstrate the success of the model in both forecast submissions submitted to the US Centers for Disease Control via the Covid19-ForecastHub 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 susceptible-exposed-infected-recovered (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 also been adopted into a Bayesian framework before, including both stochastic disease dynamics and deterministic dynamics [**?**][**?**]. Non-parametric transmissibility was included in a Bayesian SEIR model to study Ebola by Frasso and Lambert [**?**]. Time-varying transmissibility has also been studied in the frequentist setting using complex non-parametric functions [**?**]. Many efforts have been made to use SEIR models in forecasting COVID-19 [**?**][**?**] [**?**][**?**][**?**].

With the outbreak of COVID-19, accounting for testing has become a critical element in effectively using an SEIR model. Lopez et al. used a fixed case and death deviation model to model undiagnosed individuals in Spain and Italy [**?**]. Perhaps the most similar model to ours was put forth by Pei et al. [**?**]. This model was an extension to the SEIR model with time varying transmission and case and death deviation model. However, their model fixed case and death deviation model across time and was fit using Kalman Filter techniques instead of Bayesian HMC. They also used the model mostly for counterfactual scenario projections instead of forecasting. Models that have accounted for time-varying testing have mostly used the renewal style equations [**?**][**?**]. These models do not use a differential equation model as the core component, but rather parameterizes new cases as a function of the time varying reproduction number and the serial interval.

3. Data

In this analysis we use confirmed case counts and deaths as reported by the Johns Hopkins University Center for Systems Science and Engineering [**?**]. 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 [**?**], COVID-19 cases are underreported, with the fraction of all infections reported as cases for the U.S. estimated at

20-30% [**?**]. There are large discrepancies in reporting practices across the states (Figure **??**). For example, New Jersey reported an additional 1600 incident deaths when changing reporting practices to include p̈robable" deaths as well as confirmed deaths due to COVID-19 on June 25th 2020. However, many other states have at least one outlying value, usually due to backlog reporting, where a large number of deaths occurring in previous weeks are reported on a single day. We can also see from Figure **??** that some states do not report on particular days (usually weekends) leading to 0 incident deaths for the day. Some states also revise the cause of death, allowing for negative incident deaths. Some states exhibit relatively regular weekly reporting cycles, with reporting dropping off significantly on the weekends. This effect is most pronounced at the aggregate U.S. level, which shows a clear weekly cycle in reporting.

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 for incident and cumulative deaths from 1-4 weeks ahead. The quantiles used are $\mathbb{Q} = .05, .10, ..., .90, .95 \cup .01, .99$, with the median (.5 quantile) representing the point-forecast. Forecasts were made on Monday evenings and therefore use the incident data up until the Sunday before. A one-week ahead target corresponds to the following Saturday. A two-week ahead corresponds two the next Saturday and so on. Forecasts were evaluated using relative mean-absolute-error (relMAE) and the weighted-interval score (WIS) [**?**]. Relative mean absolute error is the ratio of the mean absolute error of the model of interest divided by the baseline model [**?**]. The WIS is an approximation to the continuously ranked probability score (CRPS), a commonly used metric in probabilistic forecasting [**?**][**?**].

4. Compartmental Model

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

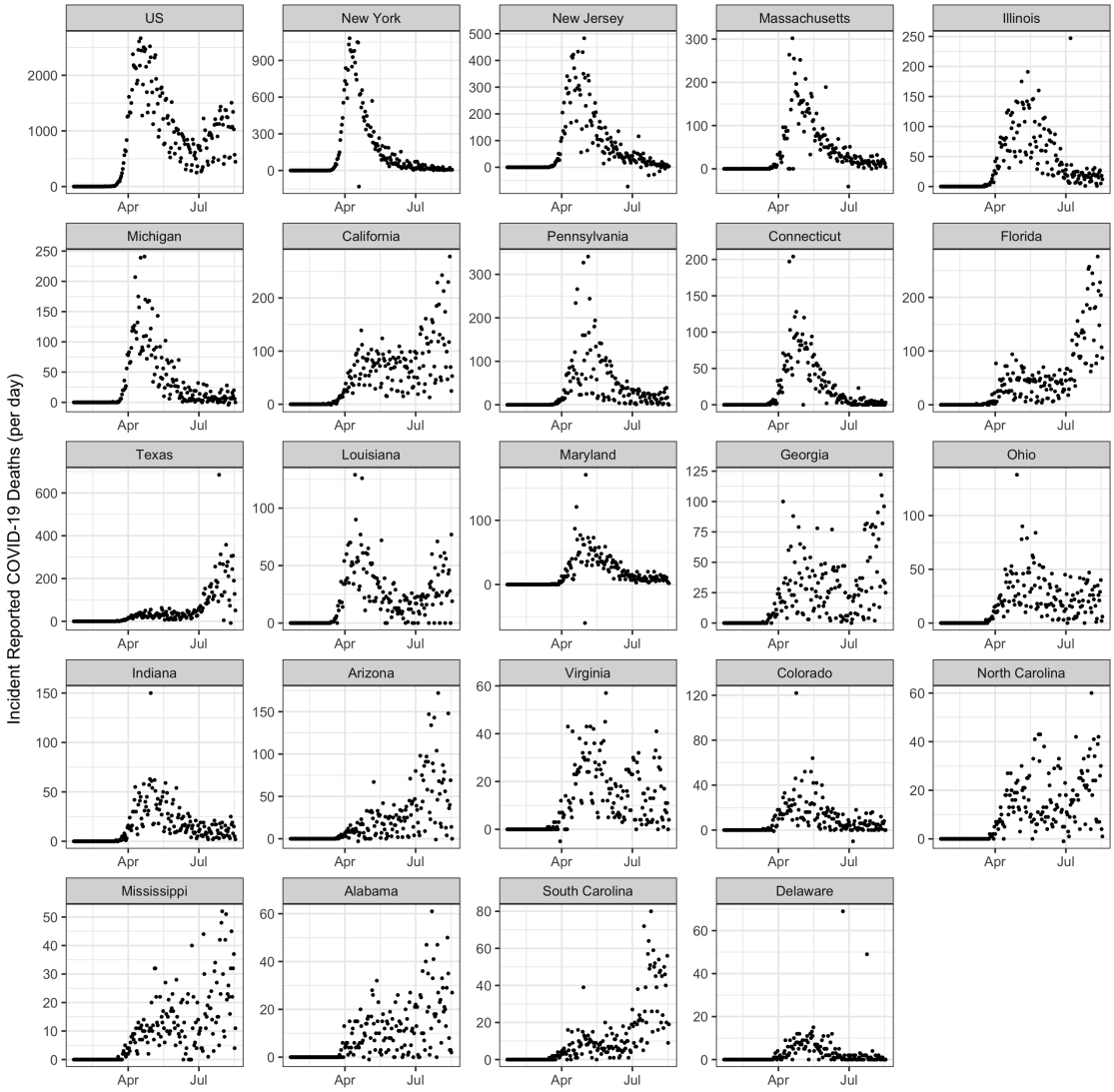


FIGURE 1. Deaths by state for any state with over 50 incident deaths for a given day. Notice the large variability in incident reporting. There are significant data-dumps, with New Jersey reporting over 1500 backlogged deaths when switching to include "probable" deaths. In some states, there appears 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.

, and deceased D_2 (Figure **??**). Here we assume everyone who is hospitalized will eventually become deceased in order to separate the rate 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 [**?**]. 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.
- 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

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

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

4.1. **Time-varying transmission parameter.** 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 [**?**]. In order to capture the aggregate effect of the interventions non-parametrically

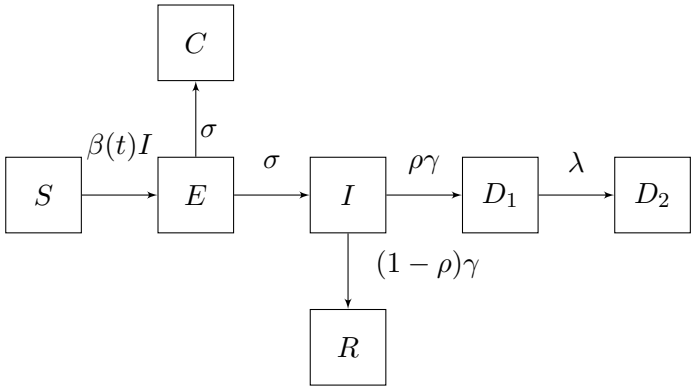


FIGURE 2. Comparmental model parameters

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 simply $\beta(t)$ at the last observed value of t.

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.

4.2. **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. Therefore, we introduce two additional parameters for the case and death deviation model of cases p_c 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 is the probability that an infected person receives a positive test result and is reported by a state or local health authority as a case. 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

$$.22, 0.38$$

. This means that we expect 30% of cases to be detected initially, as suggested by the literature [**?**]. However, we also allow this to vary by time.

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

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 concentration 100. That is, we assume that deaths due to COVID-19 are most often correctly reported [**?**].

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.

4.3. **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 [**?**]. 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 $\sigma + \gamma$ [**?**]. 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.

4.4. **Fitting.** We use the Hamiltonian Monte Carlo algorithm implemented in numpyro to fit the model to data [**?**]. 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

5. 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 Covid19-ForecastHub consortium. Second, we describe the internal evaluation used to demonstrate our model enhancements improve accuracy over a naive compartmental model.

5.1. **Real-Time Forecast Evaluation.** We began submitting forecasts to the U.S. Centers for Disease Control 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 relMAE and WIS for submissions made on 2020-05-04, 2020-05-11, 2020-05-18, 2020-05-25, 2020-06-01, 2020-06-08, 2020-06-15,2020-06-22, 2020-06-29, 2020-07-06, 2020-07-13, 2020-07-20, and 2020-07-27. Note that not all targets are observed at all weeks. This is due to 4 week ahead targets for weeks 2020-07-13 and beyond not being observable by 2020-08-01.

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 data dumps. 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.

5.2. **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 Full** Mech Bayes as of model version 3. 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 Fixed case and death deviation model** MechBayes Full with $p_{c,t}$ fixed to p_c, that is, removing the time-varying case and death deviation model.
- **MechBayes Death Only** MechBayes Full with observations on cases removed.

Note that these are nested models, with MechBayes Death Only contained in MechBayes Fixed case and death deviation model contained in MechBayes Full.

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 data dumps. 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 [**?**] [**?**][**?**] [**?**].

We can see from Figure **??** that MechBayes version 3 is able to accurately model the observed data. 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 **??** 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

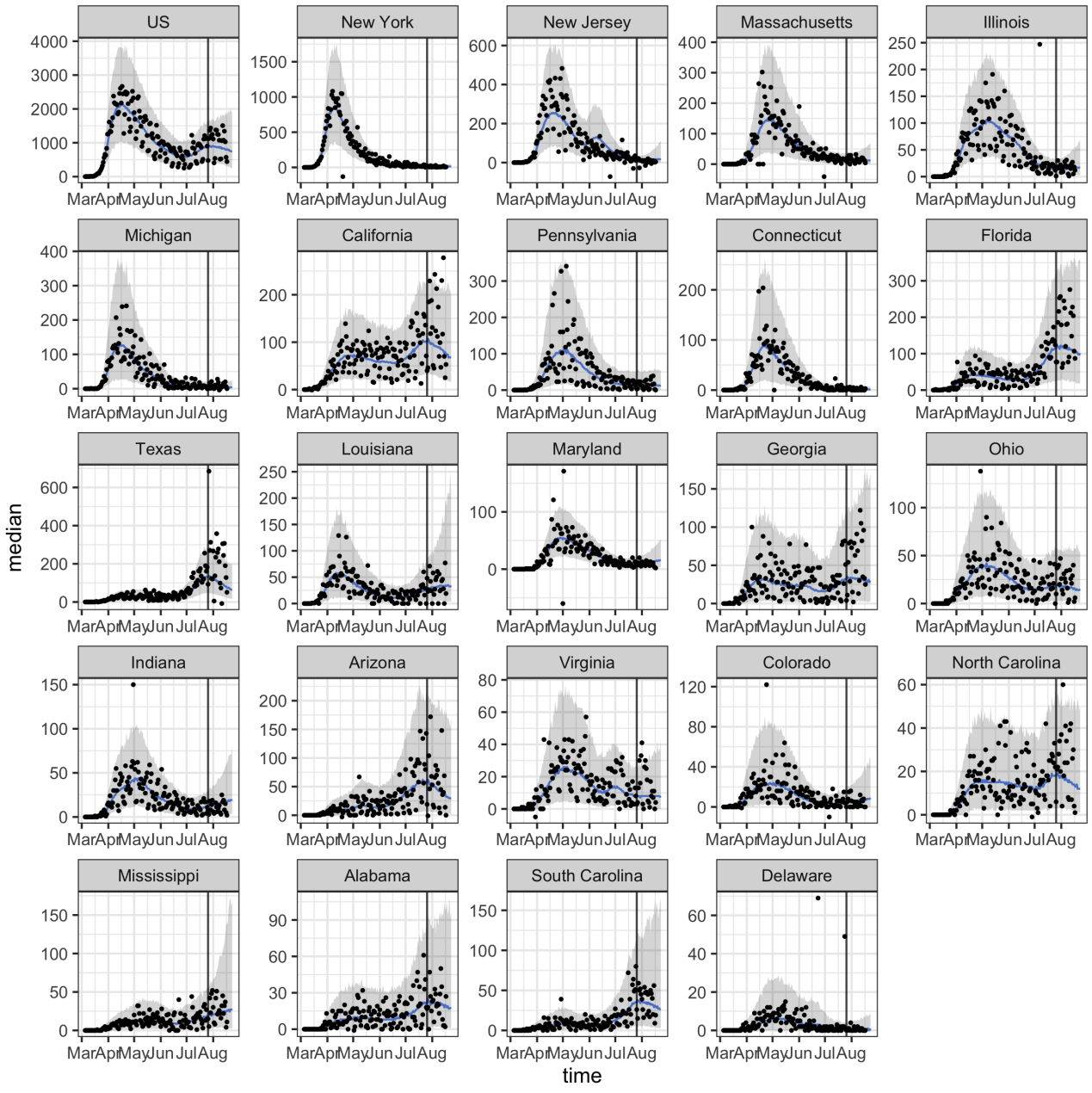
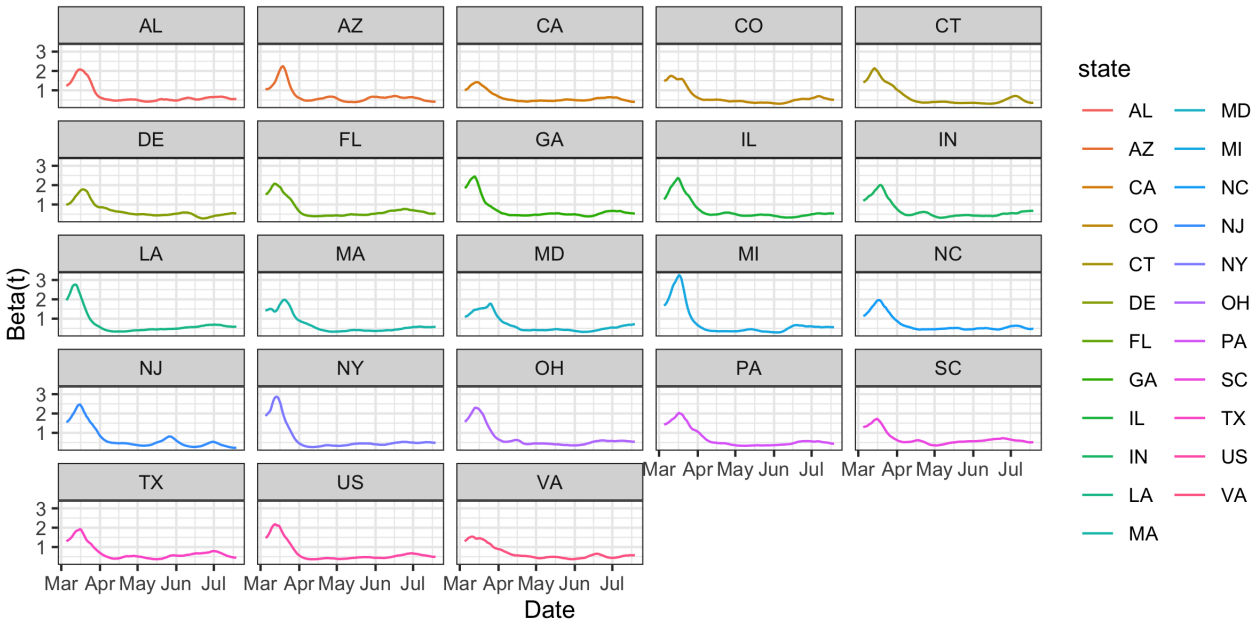
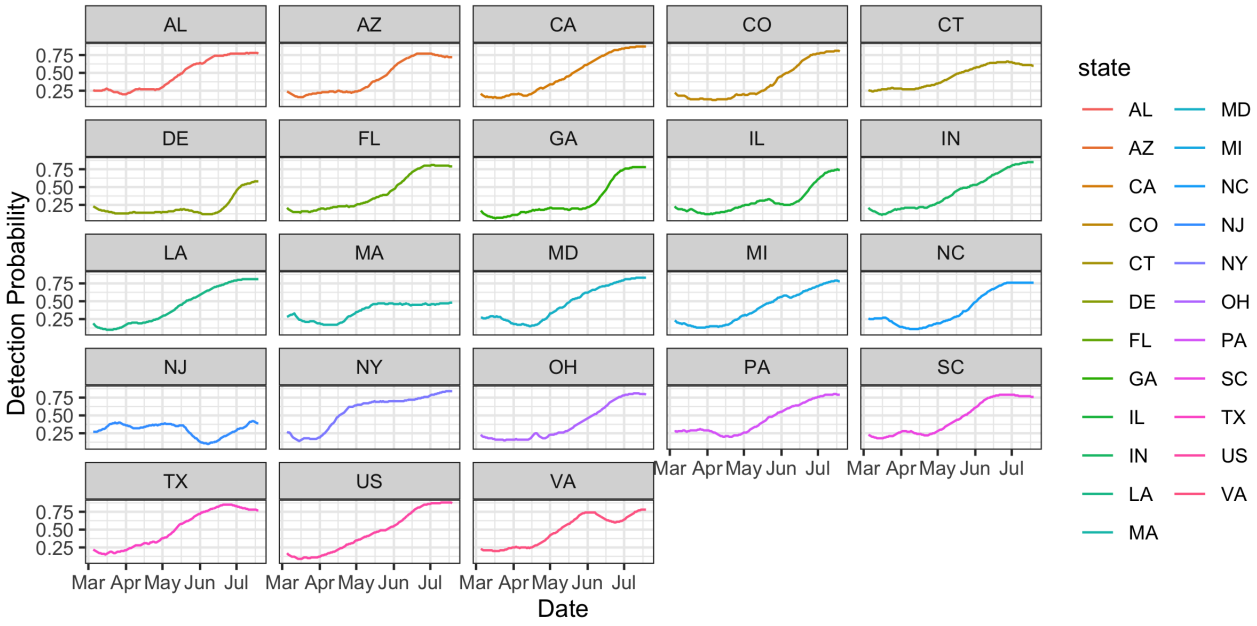


FIGURE 3. Example fit and forecast for 2020-07-26 for states with over 50 incident deaths and the U.S.. Grey bands represent 95% prediction intervals. Blue line represents median forecast. MechBayes is able to produce well calibrated fits to the data as well as accurately tracking trends in incident deaths.



(A) Time-varying beta ($\beta(t)$).



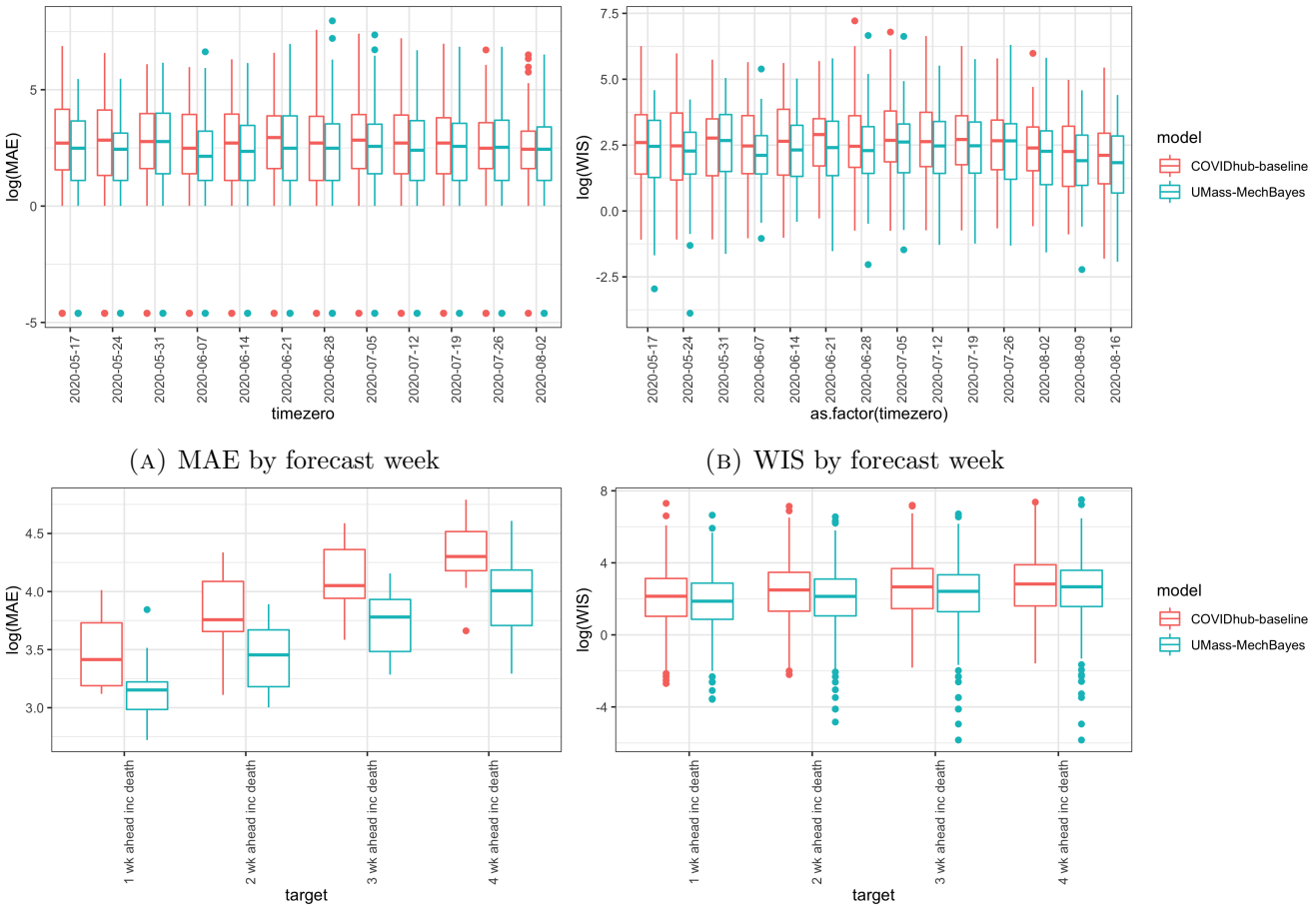
(B) Time-varying case and death deviation model ($p_{c,t}$)

FIGURE 4. A) Time varying beta parameter for all states with a day exceeding 50 incident deaths, including the U.S. by 2020-07-26. We can clearly see that the random walk is able to non-parametrically account for the non-pharmaceutical interventions retrospectively, with transmission dropping in states like New York quite rapidly. While it may seem that many states follow the same trend, this is slightly misleading since the prior pulls $\beta(t)$ upward for t close to 0 in states that did not experience community transmission in March. B) Time varying case and death deviation model for example states. The time varying detection random walk is able account for the increase in testing and other reporting anomolies. There is remarkable similarity among the estimated case and death deviation model trajectories between states.

high, with all observations outside of the 95% prediction interval falling below. Figure **??** 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 weekly reporting cycles.

We can see from Figure **??** that MechBayes is able to learn to adapt to the evolving pandemic situation. Panel A shows the time-varying transmission $\beta(t)$ for three example states, CA, FL, and NY as well as the U.S. We can see that $\beta(0)$ is centered on our prior but as data comes in, the estimate increases. This is especially true in NY, where the epidemic took off quickly in March. However, the model is then able to adjust to the varying levels of non-pharamceutical interventions present in each of the three states as well as across the U.S. This radically reduces the transmissibility parameter by the first week of April 2020. This is consistent with a peak in overall deaths two weeks later in mid-April 2020. There seems to be some estimation issues at the boundary of $\beta(t)$ where the number of cases does not match the number of deaths, since the cases have not yet converted to deaths. This results in the model underestimating transmissibility to reduce the flow through the compartments.

We can also see from Figure **??** that MechBayes is able to account for the drastic increase in testing that has occurred across the U.S. since March 2020. Panel B also shows that our prior estimate of 30% of cases being detected, may have been too high, as all regions show a dip in case and death deviation model before climbing again. Note that interpreting this strictly as time-varying detection is obscuring the fact that this parameter $p_{t,d}$ can soak-up any excess variation beyond the ability of cases and the case-fatility ratio to explain the number of deaths. That is, the time-varying case and death deviation model is able to "de-couple" cases and deaths beyond the case-fataility ratio regardless of the underlying reason (whether that be an increase in testing or shifting age distribution of cases). Thus, interpretation of this parameter as a strict mapping to testing is incorrect. As a forecasting



(A) MAE by forecast week

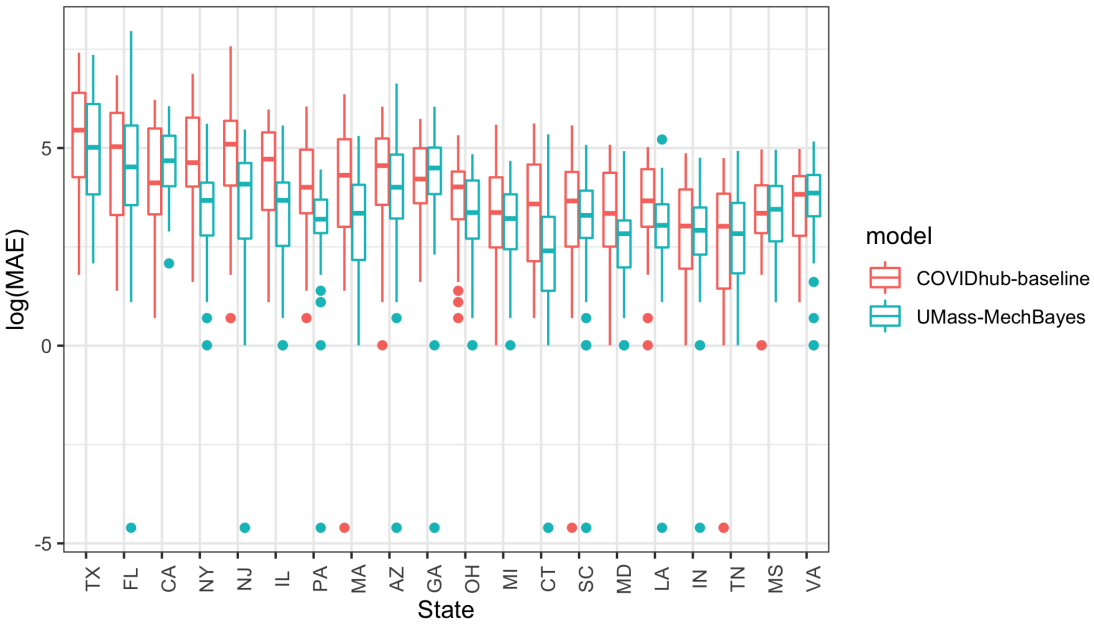
(B) WIS by forecast week

(C) MAE by target

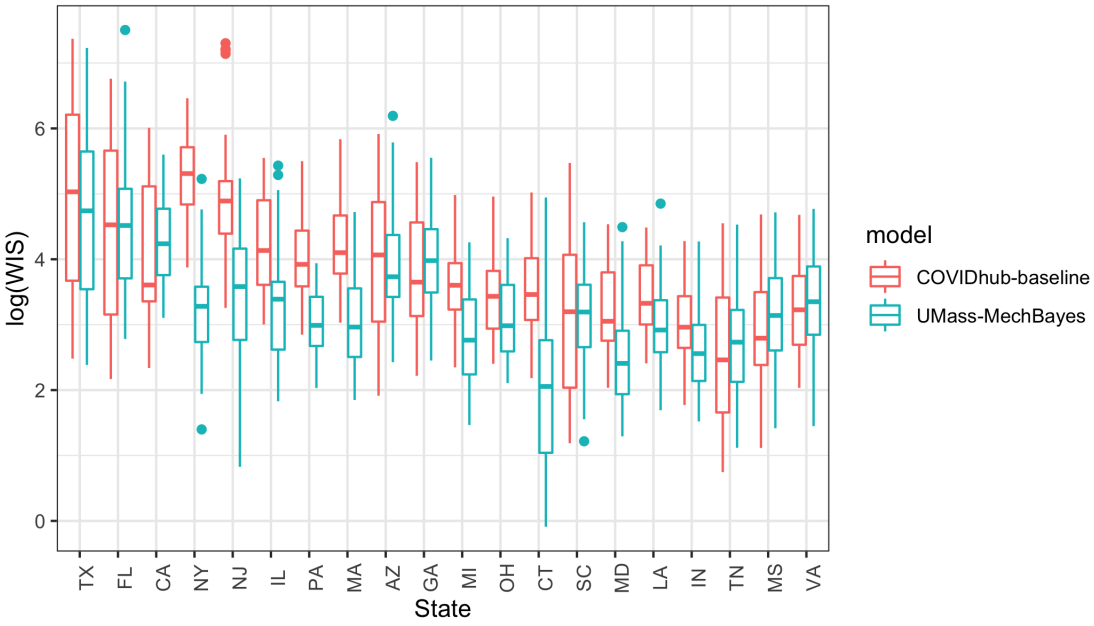
(D) WIS by target

FIGURE 5. Scores from Covid19-ForecastHub broken down by target and forecast week. We use the relative MAE metric which takes the MAE of MechBayes model and divides by the MAE of the baseline model. Here we can see that the MechBayes model improves in both MAE and WIS over time, consistently beating the baseline model in the month of July 2020. Finally, we can see that as the horizon increases from 1 to 4 weeks ahead, both relMAE and WIS increase, reflecting an increase in difficulty of forecasting further ahead in time. Note that targets averaging over forecast week include an average over all model versions.

model, we only need the ability to non-parametrically model deviations and reporting issues in cases.



(A) MAE by region



(B) WIS by region

FIGURE 6. Scores from Covid19-ForecastHub broken down by region for the 20 regions with the highest incident deaths. 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 trends are similar for WIS, with an even more pronounced improvement for states with large death counts.

6. Real-Time Model Results

We next turn to the comparison of MechBayes against the Covid19-ForecastHub 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 Covid19-ForecastHub for more details. [**?**].

We begin by breaking down the results by week the forecast was made (forecast week) and averaging over both region and target. As we can see from Figure **??** (A,B), MechBayes V1 outperformed the Covid19-ForecastHub baseline on MAE or WIS when broken down by forecast week. The same is true for MechBayes V2 with the exception of June 26th, where the Covid19-ForecastHub baseline model slightly outperformed MechBayes V2. As we can see from **??**, 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 **??** (C,D). Note that this break-down is averaged over target (1-4 week ahead), but also forecast week, meaning that we average over each of the model versions. We feel this is important, as it reflects the real-time accuracy of our evolving model efforts, instead of cherry-picking the best model. However, we still see consistent improvements in MAE under MechBayes when broken down by region. We see the largest improvements in the regions where the baseline model had the highest MAE, meaning that MechBayes improves forecasts in regions with more significant viral activity.

We break down the results by target by averaging over region and forecast week **??** (E,F) Here we can see uniform improvement over the baseline model by MechBayes in terms of MAE and WIS. 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 and WIS to demonstrate statistically significant advantages in using MechBayes over the baseline model. In order to do this, we just a random effects regression model of the form,

$$log(MAE_{m,t,r,h} + \delta) = \beta_0 + \beta 1 * h_1 + ... + \beta_3 * h_3 \tag{11}$$

$$+ \beta_4 * h_1 * mb + ...\beta_8 * h_4 * mb \tag{12}$$

$$+ b_r + \epsilon \tag{13}$$

$$b_r \sim N(0, \Sigma_b^2) \tag{14}$$

$$\epsilon \sim N(0, \sigma^2) \tag{15}$$

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,

7. Ablation Test Results

We can see from Figure **??** that MechBayes Full is consistently better than MechBayes Death Only or MechBayes Fixed case and death deviation model. The only exception seems to be June 22nd 2020 for the two week ahead target. Late June 2020 is when many states in the U.S. began to see an uptick in cases again, with Texas, Florida, and California seeing their largest total case counts since the beginning of the pandemic. Since MechBayes is conditional on the current level of interventions, forecasts made from June 22nd assumed the same level of intervention present on June 22nd. During this time, each state was

	Estimate	Std. Error	df	t value	Pr(>|t|)
β_0	1.96	0.24	56.74	8.24	0.00*
β_{h_1}	0.45	0.10	4643.00	4.66	0.00*
β_{h_2}	0.61	0.10	4643.00	6.33	0.00*
β_{h_3}	0.75	0.10	4643.00	7.63	0.00*
$\beta_{h_1 mb}$	-0.18	0.10	4643.00	-1.88	0.06
$\beta_{h_2 mb}$	-0.37	0.10	4643.00	-3.85	0.00*
$\beta_{h_3 mb}$	-0.27	0.10	4643.00	-2.78	0.01*
$\beta_{h_4 mb}$	-0.37	0.10	4643.00	-3.68	0.00*

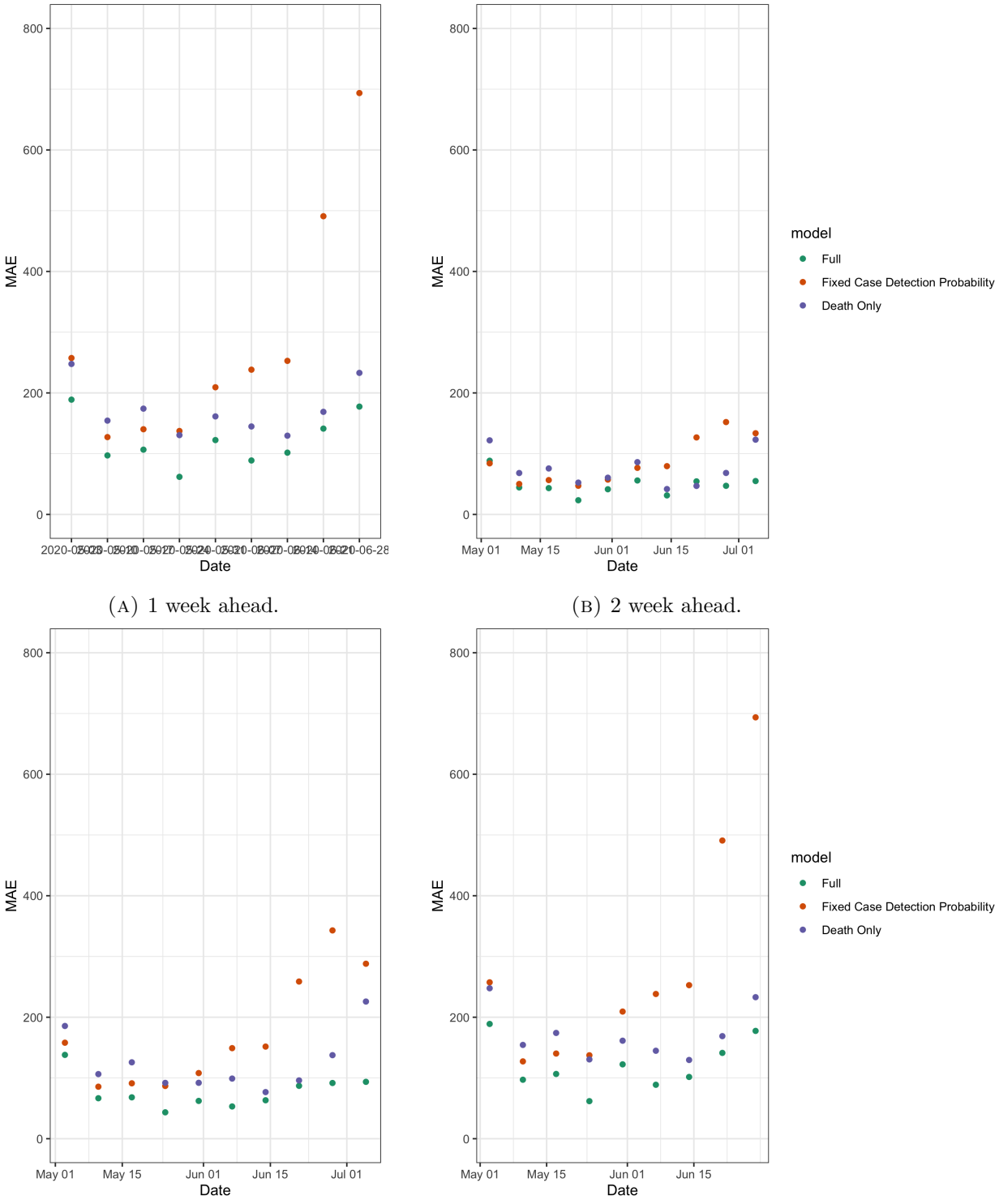
TABLE 1. Coefficient estimates and t-values for MAE evaluation model. We can see that MechBayes performs statistically significantly better than the baseline model for 2-4 weeks ahead. The performance increase does not seem to follow a particular pattern.

re-opening various establishments, while cases were rising. MechBayes Full forecasted an exponential growth in cases, which was reflected in a large over prediction two weeks later on July 4th. Finally, we can also see that the difference in MAE between MechBayes Full and the competing models increases as forecast horizon increases. This suggest that MechBayes Full is a better long term forecasting model.

We can also see that the MechBayes Death Only is consistently better than MechBayes Fixed case and death deviation model. This may be evidence that naively including case data, without adjusting for time-varying testing rates, may be worse than not including it at all.

8. 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 [**?**]. Our experiments led us to the following conclusions about the performance of this model and the underlying methodology.



(A) 1 week ahead.

(B) 2 week ahead.

(C) 3 week ahead.

(D) 4 week ahead.

FIGURE 7. Ablation test MAE broken down by forecast week and target. We can see that MechBayes Full performs better than MechBayes No Detection Random Walk and MechBayes Only death in almost all breakdowns. We can also see that the improvement becomes more pronounced at larger horizons, suggesting that MechBayes Full is a better longer term forecasting model.

- **Adding case data when predicting deaths is helpful but only when accounting for data quality issues**. Our ablation tests (Figure **??**) 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 time-varying case detection rate 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 case and death deviation model.
- **MechBayes is accurate when compared to a baseline model**. As Figure **??** 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 (95%) are almost completely separated. These results are statistically significant for 2-4 weeks ahead, suggesting that the biggest improvements from MechByes over the baseline model comes in longer term forecasting. Additionally, as seen in Figure **??** , the biggest gains in performance, in both MAE and WIS, occur in regions with the largest total deaths counts. This is a desirable feature of a pandemic forecasting model.
- **MechBayes is biased high**. We can see from Figure **??** that MechBayes was biased high with respect to the median forecast. We suspect this is due to the inherent exponential growth of an SEIR model when the number of susceptibles is small. Since overall prevalence in any state is well below the number needed to reach herd immunity, when the model sees an increase in cases it treats this as exponential growth, which translates into exponential growth in deaths. However, in the recent forecasts the bias is consistently lower than the Covid19-ForecastHub baseline model. Note the extreme low bias in the first submission is due to the omission of the U.S. in the submission. **is this true**

- **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. 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.
- **MechBayes is probabilistically well-calibrated compared with the baseline model**. Small WIS values are better in terms of calibration [**?**]. The two sources of uncertainty in MechBayes come from the distribution over the differential equation parameters and the observation noise, unlike a fully stochastic model that has inherent variability within the differential equation transitions. This suggest that even a deterministic model core can produce well calibrated prediction intervals by capturing the uncertainty in parameter estimation and observation noise.
- **Allowing for time-varying transmissibility is necessary to non-parametrically capture the effect of non-pharamceutical 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,

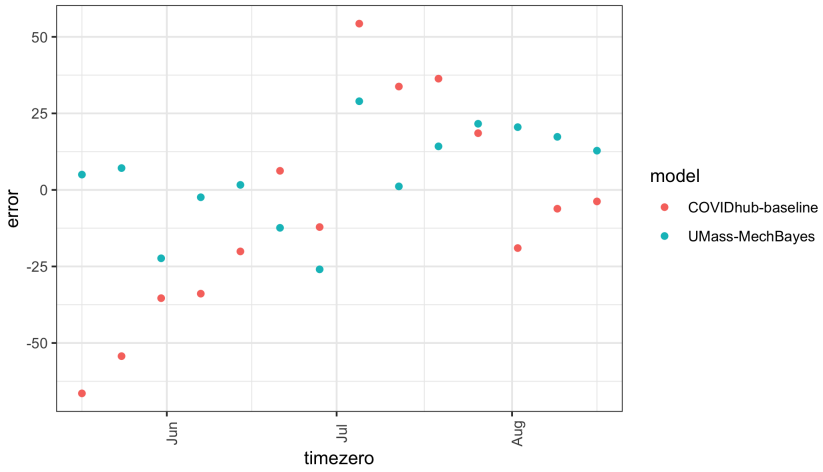


FIGURE 8. Bias of MechBayes and COVID-Baseline as a function of time.

then $\beta(t)$ could vary enough to match the data exactly. This would clearly attribute reporting issues as true changes in transmissibility. Bypassing the epidemiological interpretation that compartmental models provide.

- **MechBayes Full outperforms the other versions on probabilistic scoring measures**. For simplicity, we only included MAE results for the ablation test. However, as can be seen from Figure **??**, MAE and WIS are highly correlated. We therefore conclude that MechBayes Full is also better calibrated than the other two competing models.

9. 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 the Covid19-ForecastHub 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.

10. Appendix

10.1. **A1.**

10.2. **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.

$$E_0 \sim \text{Unif}(0, 0.02N)$$

$$I_0 \sim \text{Unif}(0, 0.02N)$$

$$D_{1_0} \sim \text{Unif}(0, 0.02N)$$

$$D_{2_0} \sim \text{Unif}(0, 0.02N)$$

$$R_0 \sim \text{Unif}(0, 0.02N)$$

This allows us to initialize the process model:

(16)

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

10.3. **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) [**?**]. 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 [**?**]. 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.

10.4. **A2.** The Covid19-ForecastHub began soliciting forecasts in the beginning of April 2020 for 1-4 week ahead cumulative deaths. We began submitting the first version on April 20th 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. Our model went through three distinct iterations as we evaluated performance in real-time.

- **Version 1: April 20th 2020 through May 10th 2020**. Model had normal observation noise (instead of negative binomial) and a non-time varying case and death deviation model. Model was fit to cumulative deaths and cases.

$$\text{Cases}_t \sim N(p_c * C_t, \sigma_c^2) \quad \text{Deaths}_t \sim N(p_d * D_{2_t}, \sigma_d^2) \tag{17}$$

- **Version 2: May 10th 2020 through May 24th 2020**. Model had normal observation noise (instead of negative binomial) and time varying case and death deviation model. Model was fit to cumulative deaths and cases.

$$\text{Cases}_t \sim N(p_{t,c} * C_t, \sigma_c^2) \quad \text{Deaths}_t \sim N(p_d * D_{2_t}, \sigma_d^2) \tag{18}$$

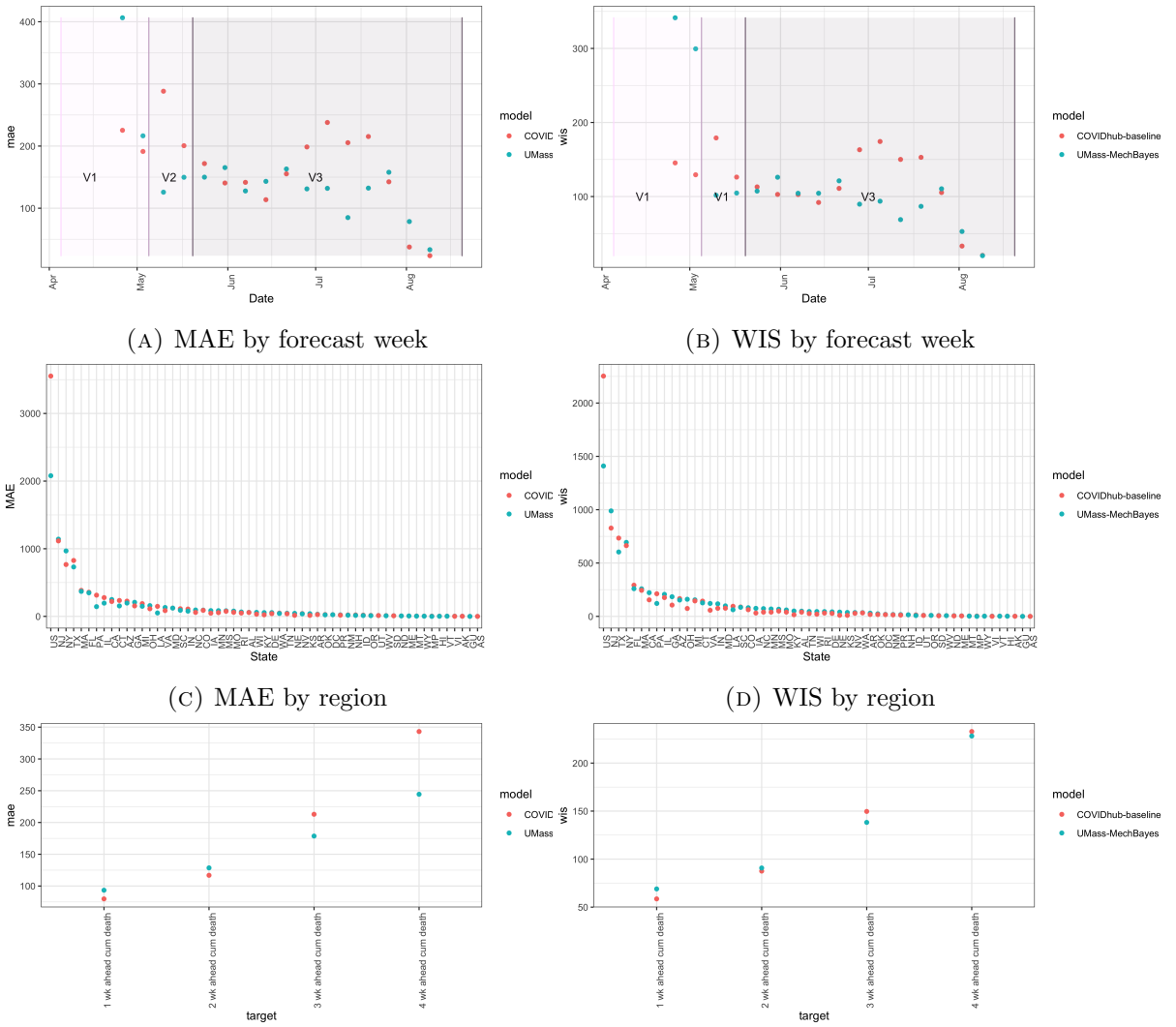
- **Version 3: May 24th 2020 through August 1st 2020**. Model had negative binomial observation noise and time varying case and death deviation model. Model was fit to incident deaths and cases.

$$\text{Cases}_t \sim NB(p_{c,t} * [C_t - C_{t-1}], \sigma_c^2) \quad \text{Deaths}_t \sim NB(p_d * [D_{2_t} - D_{2_{t-1}}], \sigma_d^2) \tag{19}$$

These three versions highlight the complexities of forecasting during a real-time pandemic. Models evolve due to real-time evaluations by responding to the unique data collection environments of each pandemic. We use the model submissions made in real-time, under the corresponding version of the model as dated above, evaluated on both MAE and WIS for the weeks of 2020-05-05, 2020-05-10, 2020-05-17, 2020-05-24 2020-05-31, 2020-06-07, 2020-06-14, 2020-06-2, 2020-06-28, 2020-07-05, 2020-07-12, 2020-07-19, and 2020-07-26.

Note that not all targets are observed at all weeks. This is due to 4 week ahead targets for weeks 2020-07-12 and beyond not being observable by 2020-08-01.

These two versions highlight the complexities of forecasting during a real-time pandemic. Models evolve due to real-time evaluations by responding to the unique data collection environments of each pandemic.



(A) MAE by forecast week

(B) WIS by forecast week

(C) MAE by region

(D) WIS by region

(E) MAE by target

(F) WIS by target

11. Real-Time Model Results

Figure 9. Scores from Covid19-ForecastHub broken down by region, target and forecast week. Here we can see that the MechBayes model improves in both MAE and WIS over time, consistently beating the baseline model in the month of July 2020. We can also see that MAE varies heavily by region, which is an artifact of both population size and number of covid deaths. However, in regions with large MAE (left side) we see a significant improvement over the baseline model in terms of both MAE and WIS. Finally, we can see that as the horizon increases from 1 to 4 weeks ahead, both MAE and WIS increase, reflecting an increase in difficulty of forecasting further ahead in time. Note that targets averaging over forecast week include an average over all model versions.