MECHANISTIC BAYESIAN FORECASTS OF COVID19

GRAHAM GIBSON, NICHOLAS REICH, DAN SHELDON

Abstract

As of August 1st 2020 the Covid-19 pandemic has caused over 4 million infections and over 150,000 deaths in the United States alone. 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, backlog reporting, and limited testing. Classical infectious disease differential equation models, such as the susceptible-exposed-infected-recovered model, do not take these data issues into account. We propose a novel Bayesian compartmental model that allows for non-parametric modeling of non-pharamacuetical interventions, time-varying testing, and joint observations on case counts and deaths. The model has been used to submit forecasts to the Covid-hub repository organized by the Reich Lab. We examine the performance relative to the baseline model in addition to performing a nested model comparison of our extensions to the classic models. We demonstrate a significant gain in both MAE and probabilistic scoring measures when taking into account data quality issues during a real-time pandemic.

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. As of July 2020, there are over 4 million confirmed infections and over 150,000 deaths due to COVID-19 [1]. Understanding the future trajectory of the pandemic is crucial for minimizing the impact across the

nation in terms of healthcare burden, economic recession, and political stability. Forecasts of incident and cumulative deaths due to COVID may help in resource allocation, vaccine clinical trial planning, and re-opening strategies. Along with non-pharamacutical interventions, forecasts are one of the few public health tools available to help fight the pandemic. Infectious disease forecasting has been demonstrated to benefit public health decision makers during annual influenza outbreaks [2]. 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. The COVID-19 pandemic has seen a resurgence in 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 the susceptible, exposed, infected, and recovered compartment [3]. The model is specified by setting the rates of flow 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. Compartmental models have been used to effectively model and forecast disease in non-pandemic situations. These include complex compartmental models for Influenza forecasting [4][5][6], including a retrospective model evaluation of the 1918 Influenza pandemic [7]. Compartmental models have been used not just in respiratory disease but in Ebola [8], Measles [9], Dengue [10] and a wide variety of other disease.

Compartmental models have also been adopted into a Bayesian framework before, including both stochastic disease dynamics and deterministic dynamics [11][12]. In fact, non-parametric transmissibility was included in a Bayesian SEIR model to study Ebola by Frasso and Lambert [13]. Time-varying transmissibility has also been studied in the frequentist setting [14]. With the outbreak of COVID-19, accounting for testing has become a

critical element in using an SEIR model. Lopez et al. used a fixed detection probability to model undiagnosed individuals in Spain and Italy [15]. Many efforts have been made to use SEIR models in forecasting COVID-19 [16][17] [18][19][20]. Perhaps the most similar model was put forth by Pei et al. [21]. This model was an extension to the SEIR model with time varying transmission and detection probability. However, their model fixed detection probability 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. Another relatively similar model was put forth by Abbot et al. [22]. This model does use a time-varying transmission rate and detection rate, but does not use a differential equation model as the core component, but rather parameterizes new cases as a function of the time varying reproduction number.

Emerging pandemics create a unique set of challenges for accurately predicting future deaths. These include, but are not limited to, severe under reporting of cases due to asymptomatic transmission, time-varying testing rates, and both the addition and removal of control measures such as social distancing, lockdown, and mask use. In this work, we introduce a set of extensions to the classic compartmental model that are able to account for the real-world and real-time complexities of infectious disease forecasting during a pandemic. We demonstrate the success of the model in both real-time forecast submissions as well as an ablation test to demonstrate the additional forecast accuracy of our extensions. 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. Data

In this analysis we use confirmed case counts and deaths as reported by the Johns Hopkins University Center for Systems Science and Engineering [1]. This a time series dataset which we truncate to begin March 1st 2020 to April 27th 2020 and captures all 50 states, as well as Guam, Puerto Rico, and American Samoa. As noted in [23], COVID-19 cases are often dramatically under-reported, with reporting rates for the U.S. estimated at 20-30% [24]. In addition, there has also been severe temporal variation in the percent of symptomatic cases [24]. There are large discrepancies in reporting practices across the states (Figure 1). Most notably, New Jersey reported an additional 1600 incident deaths when changing reporting practices to include "probable" deaths as well as confirmed deaths due to COVID-19. However, many other states have at least one outlying value, usually due to backlog reporting, where a large quantity of previous deaths is reported on a single day. We can also see from Figure 1 that some states do not report on particular days (usually weekends) leading to 0 incident deaths for the day. We can also see relatively regular weekly reporting cycles, with reporting dropping off significantly on the weekends. Finally, we can see that while some states have reached their first peak and dropped off, others are starting to see a rapid increase in confirmed deaths, such as the Carolinas, Texas, and Florida.

In order to organize collaborative efforts across the modeling community, the ReichLab of UMass Amherst has developed the COVID-HUB forecast repository and visualization tool [**?**]. The COVID-HUB team have been soliciting forecasts for 1-4 week ahead incident and cumulative deaths since the beginning of the April 2020. A COVID-HUB forecast is represented by a set of quantiles for incident and cumulative deaths from 1-4 weeks ahead. The quantiles used are $\mathbb{Q} = .05, .10, ..., .90, .95 \cup .01, .99$

Forecasts are due 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 are stored in the generic repository Zoltar, which allows teams to access both their model and a baseline model [25]. Zoltar also scores models according to mean-absolute-error (MAE) and the weighted-interval score (WIS) [26]. This allows forecasts to be evaluated both point-wise and probabilistically. We can think of WIS as an approximation the the log-score, a commonly used metric in probabilistic forecasting [**?**][**?**].

3. Compartmental Model

In a given time-step (e.g. one day), each member of the population of interest belongs is assumed to be in one of the mutually-and-exhaustive compartments: Susceptible S, Exposed but not yet infectious E, Infectious I, Recovered R, hospitalized before death D_1 , and finally deceased D_2. 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. 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 latent period
- γ: rate of transition from the infectious state I to no longer being infectious; i.e., $1/\gamma$ is the expected duration of the infectious period
- ρ: fatality rate
- 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)

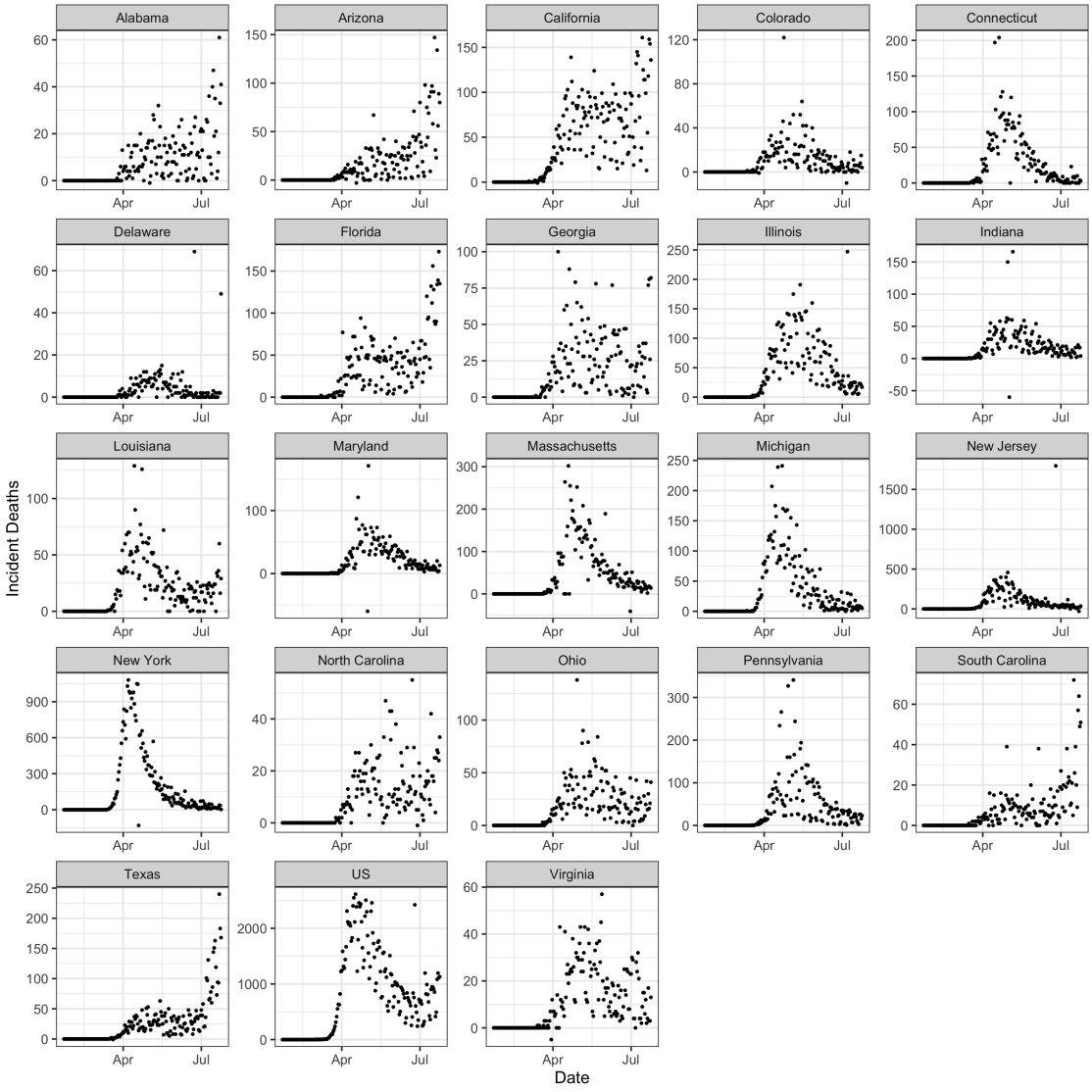


FIGURE 1. Deaths by state for any state with over 50 incident deaths for a given day. Notice the large variability in incident reporting. There are significant data-dumps, with New Jersey reporting over 1500 backlogged deaths when switching to include "probable" deaths. In some states, there appears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

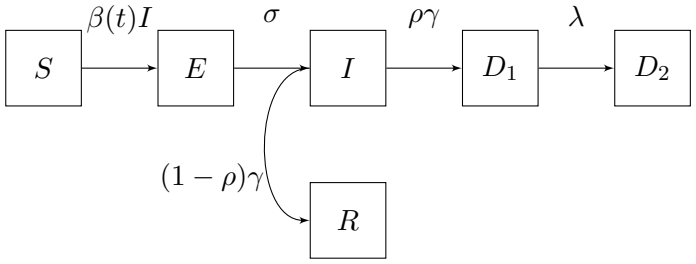


FIGURE 2. Comparmental model parameters

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

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

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

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

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

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

, where RK4 is the Runge-Katta 4th order approximation (see numpyro ode docs) [**?**].

3.1. **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 we choose a flexible model for the time-varying transmission parameter. We allow $\beta(t)$ to vary as follows,

$$log(\beta(t)) \sim N(log(\beta(t-1), \sigma_\beta^2)$$

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 β_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 I. To avoid instability issues, especially when forecasting, we set the random walk value for forecasting to the average of the last 10 days. This is a large enough window to smooth over most reporting issues (excepting data dumps).

3.2. **Observation Model.** The observed data used to fit the model is based on time-series data of confirmed cases $Cases_t$ and recorded deaths $Deaths_t$. For a given state and day, the change in the confirmed cases and reported deaths are subset of the number of infections $I(t)$ and underlying number of deaths $D2(t)$, respectively. Therefore, we introduce two additional parameters for the detection probability of cases p_c and the detection probability 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 infectious person receives a positive test result and is confirmed as case. We assume its prior distribution is given by $p \sim \text{Beta}(15, 35)$,

such that $\mathbb{E}[p_c] = 0.3$ with concentration 50. This means that we expect 30% of cases to be detected initially. However, we also allow this to vary by time.

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

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.

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

$$\text{Cases}_t \sim NB(p_{c,t} * I_t, \sigma_c^2) \tag{2}$$

$$\text{Deaths}_t \sim NB(p_d * D_{2_t}, \sigma_d^2) \tag{3}$$

3.3. **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 and exposed people at the start of the epidemic.

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

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

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

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

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

This allows us to initialize the process model:

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

3.4. **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) [cite]. 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 (cite). 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 concentration of 100. Finally, our prior on the rate at which dying patients succumb satisfies $\lambda\ \mathbb{E}[\lambda] = 0.1$ with shape 10 corresponding to roughly 10 days in the D_1 compartment.

The identifiability of model parameters in compartmental models where the data consists of only 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 [27]. 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 $\sigma + \gamma$ [27]. 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. **I want to say a little more**

3.5. **Fitting.** We use the hamiltonian monte carlo algorithm implemented in numpyro (some citation here) to fit the model to data. That is, given a time series of confirmed cases ($C_{1:t}$) and confirmed deaths ($D_{1:t}$) we use Bayesian inference (via HMC) to obtain

$$f(\boldsymbol{\theta}|C_{1:t}, D_{1:t}) \propto f(C_{1:t}, D_{1:t}|\boldsymbol{\theta})f(\boldsymbol{\theta})$$

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

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

We use the numpyro probabilistic framework to fit the model.

4. 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 COVID-HUB consortium. Second, we describe the internal evaluation used to demonstrate our model enhancements improve accuracy over a naive compartmental model.

4.1. **COVID-HUB.** The COVID-HUB 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 detection probability. Model was fit to cumulative deaths.

- **Version 2: May 10th 2020 through May 24th 2020**. Model had normal observation noise (instead of negative binomial) and time varying case detection probability. Model was fit to cumulative deaths.
- **Version 3: May 24th 2020 through August 1st 2020**. Model had negative binomial observation noise and time varying case detection probability. Model was fit to incident deaths.

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.

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

- Identifying outliers in recently reported incident cases.
- Search for documented evidence of a data dump. These are usually recorded on state department of health websites and sometimes local news outlets.
- Manually redistribute 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 1600 daily deaths as it switched from reporting only confirmed deaths from COVID-19, to

confirmed and probable. This would have caused a drastic increase in predictions if not properly identified as data dump.

4.2. **Ablation Test.** While real-time model evaluation is valuable for understanding evolving model performance, we also perform a retrospective evaluation using three model variants. 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 Detection Probability** MechBayes Full with $p_{c,t}$ fixed to p_c, that is, removing the time-varying detection probability.
- **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 Detection Probability contained in MechBayes Full. This allows to perform a nested model comparison using MAE as the scoring metric.

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.

5. COVID-Hub Results

We can see from Figure 3 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 3 shows prediction intervals at the 95% level. However, we can also see that in some states, such as California and Florida, the model is biased high, with all observations outside of the 95% prediction interval

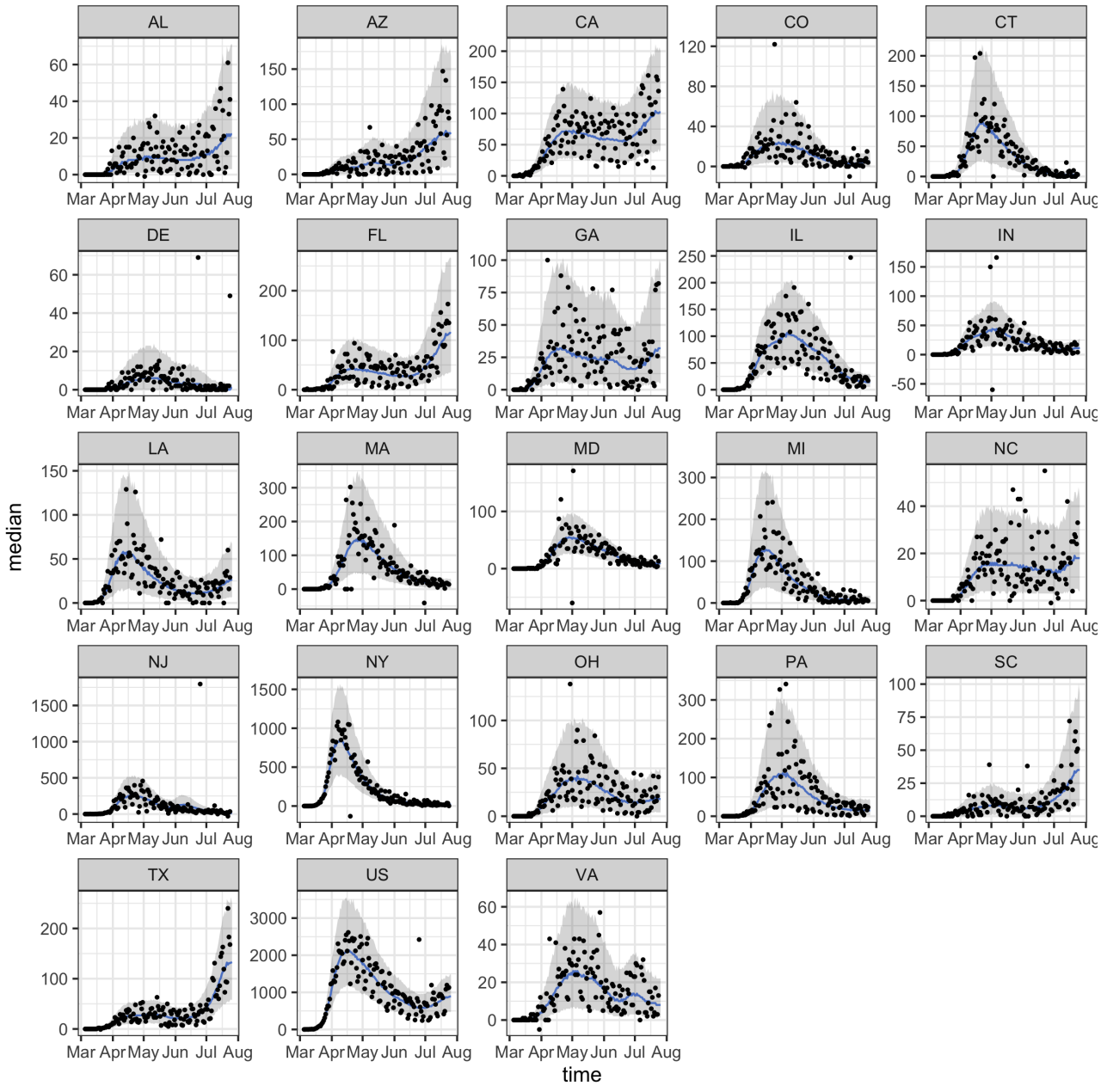
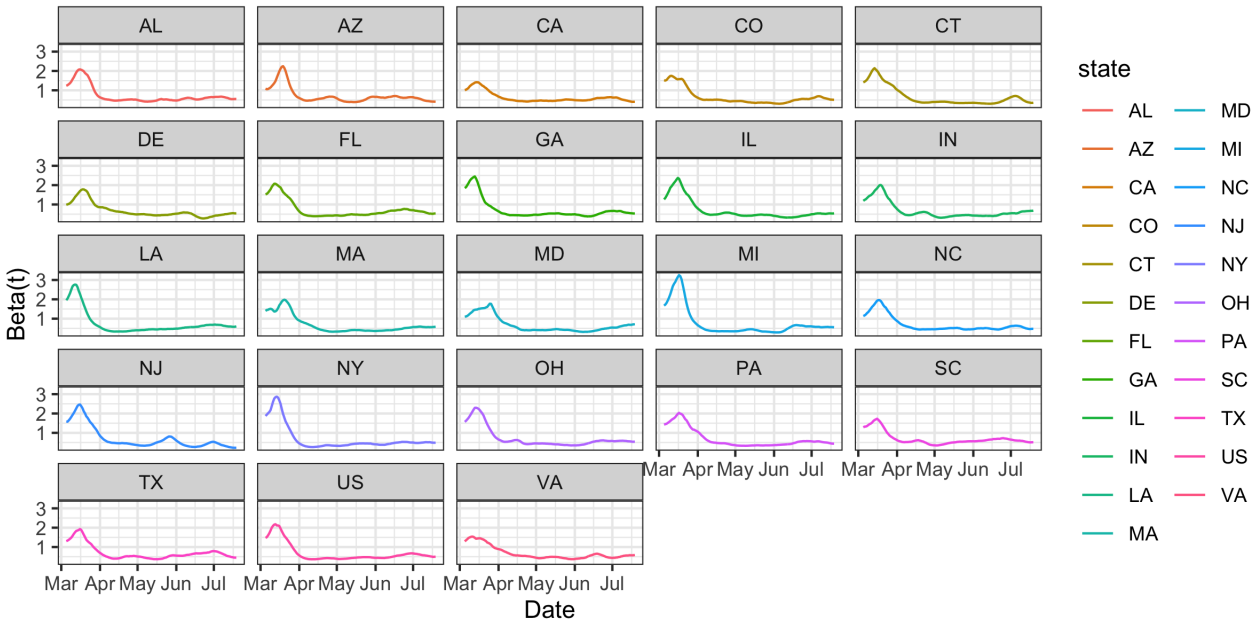
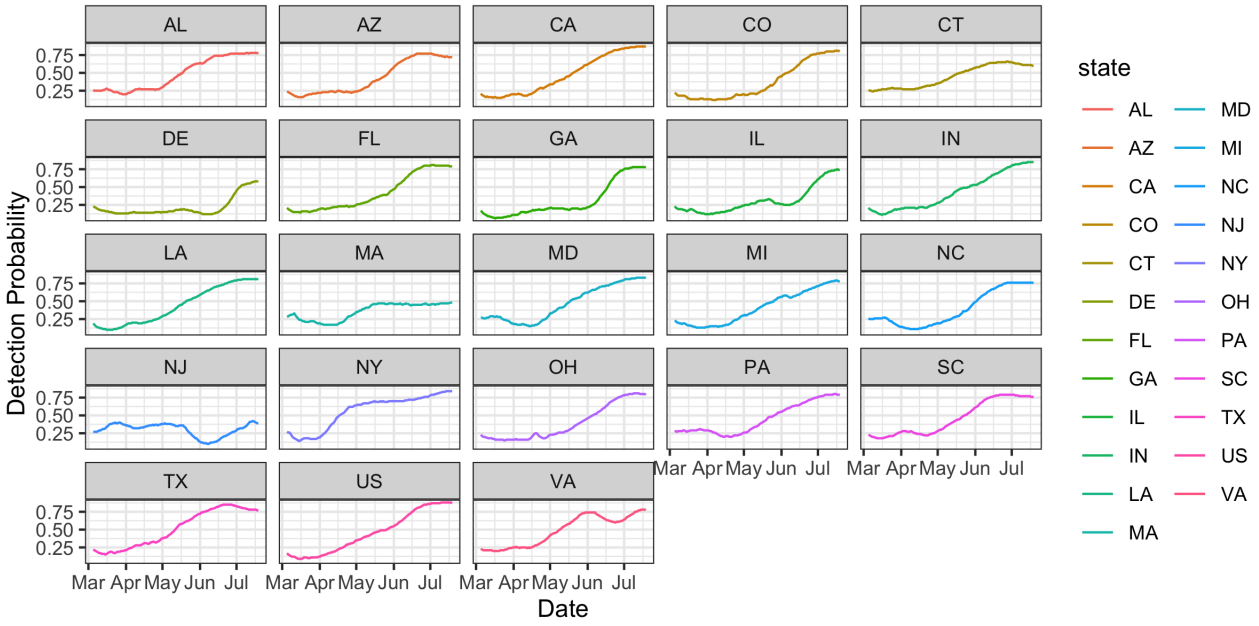


FIGURE 3. Example fit and forecast for states with over 50 incident deaths. Grey bands represent 95% prediction intervals. Blue line represents median forecast. MechBayes is able to produce well calibrated fits to the data.

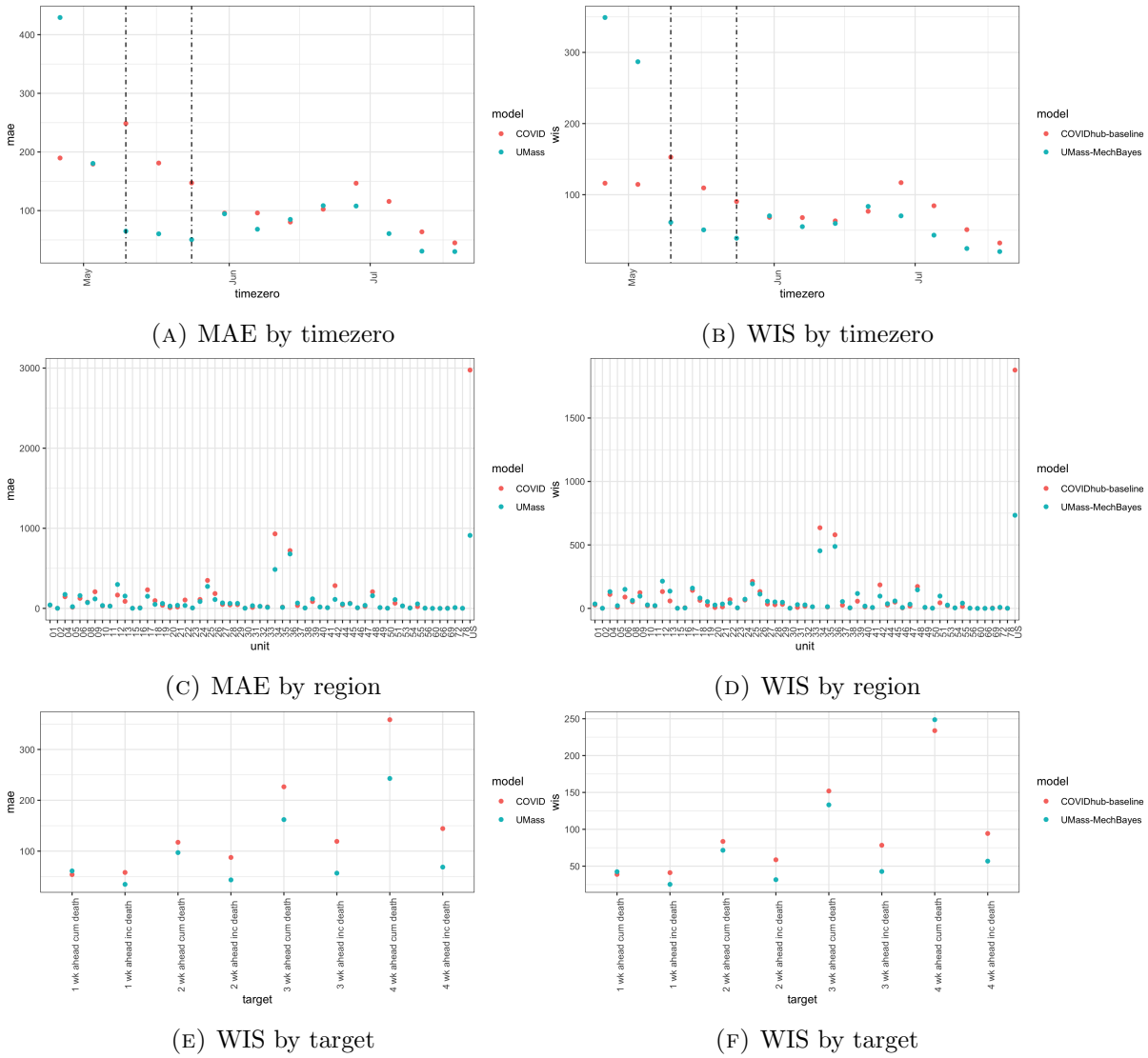


(A) Time-varying beta ($\beta(t)$).



(B) Time-varying detection probability ($p_{c,t}$)

FIGURE 4. A) Time varying beta parameter for all geographies with a day exceeding 50 incident deaths. We can clearly see that the random walk is able to non-parametrically account for the non-pharmaceutical interventions. 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 detection probability for example states. The time varying detection random walk is able account for the increase in testing. There is remarkable similarity among the estimated detection probability trajectories between states.



(A) MAE by timezero

(B) WIS by timezero

(C) MAE by region

(D) WIS by region

(E) WIS by target

(F) WIS by target

FIGURE 5. Scores from COVID-HUB broken down by region, target and timezero. 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. Note that targets averaging over timezero include an average over all model versions.

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 weekly reporting cycles.

We can see from Figure 4 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. **Need to think on this a bit**

We can also see from Figure 4 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 detection probability 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 detection probability 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 model, we only need the ability to non-parametrically model deviations and reporting issues in cases.

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

We begin by breaking down the results by week the forecast was made (timezero) and averaging over both region and target. As we can see from Figure 5 (A,B), MechBayes V1 did not outperform the COVID-HUB baseline on MAE or WIS when broken down by timezero. However, MechBayes V2 (the introduction of the time-varying detection probability), outperformed the COVID-HUB baseline model on both MAE and WIS when broken down by timezero. The same is true for MechBayes V3 with the exception of June 26th, where the COVID-HUB baseline model slightly outperformed MechBayes V3. However, as seen in Figure 1, overall incident deaths were at their lowest across the U.S. in late June and remarkably stable, meaning this week was the hardest of the weeks to beat the baseline model.

We also break down the results by geographical region, as seen in Figure 5 (C,D). Note that this break-down is averaged over target, but also timezero, 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 consistent improvements in MAE under MechBayes when broken down by region, with the exception of New York. This is mainly due to MechBayes V1 large MAE for New York in March, where most of the cases (and therefore contribution to MAE) occurred. The WIS results are similar.

Finally, we break down the results by target by averaging over region and timezero 5 (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 see the model is slightly better calibrated on incident deaths than cumulative.

6. Nested Model Comaprison

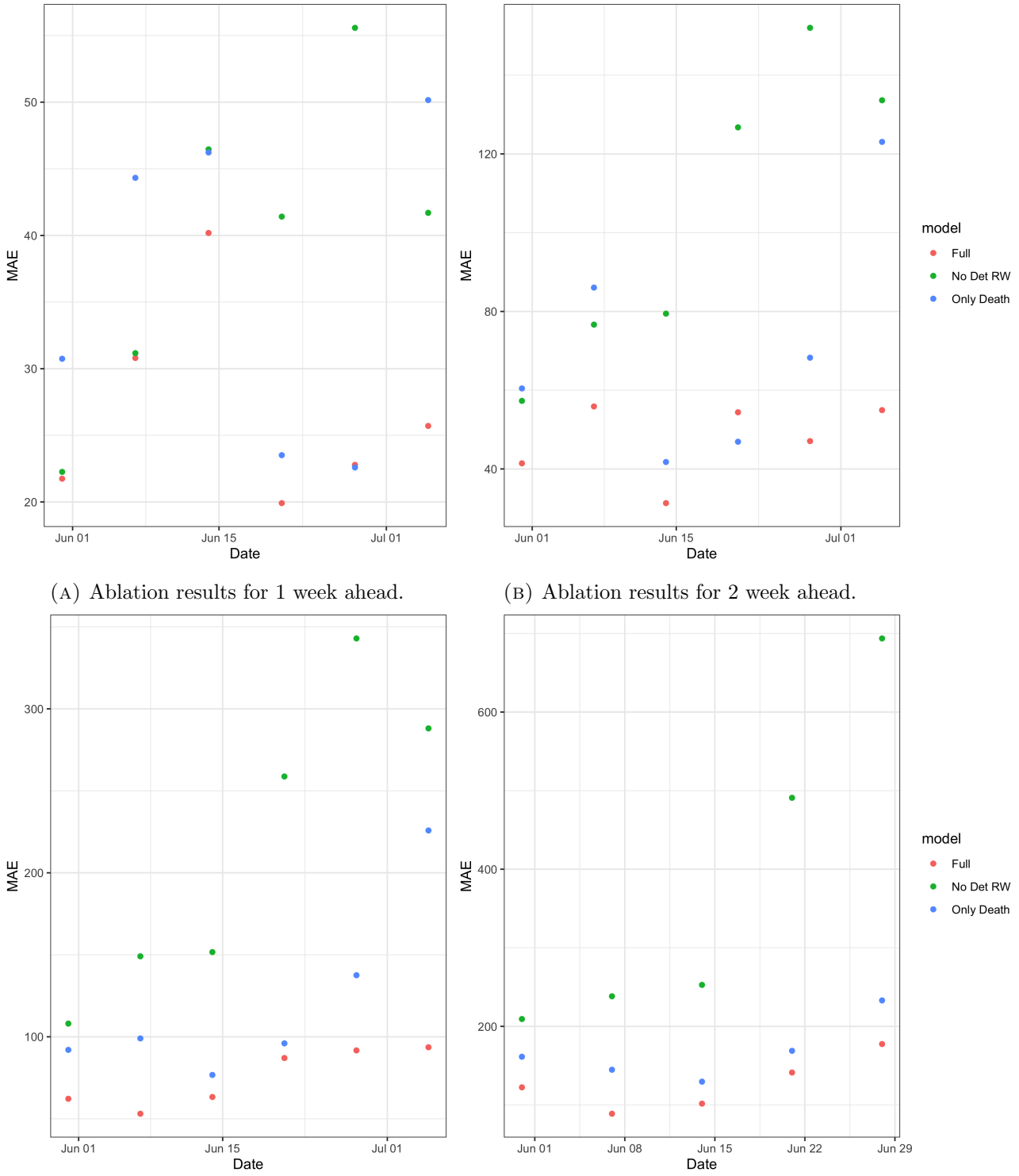
We can see from 6 that MechBayes Full is consistently better than MechBayes Death Only or MechBayes Fixed Case Detection Probability. The only exception seems to be June 22nd 2020 for the two week ahead target. Late June 2020 is when many states in the U.S. began to see an uptick in cases again, with 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 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.

We can also see that the MechBayes Death Only is consistently better than MechBayes Fixed Case Detection Probability. 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.

7. Discussion

MechBayes is a fast, fully Bayesian compartmental model capable of accounting for real-world modeling challenges during a pandemic. Our experiments led us to the following conclusions.

- **Adding case data when predicting deaths is helpful but only when accounting for data quality issues**. Our ablation test (Figure 6) clearly shows that time-varying detection probability is a key feature in the model for reducing MAE of forecasts. As we can see in Figure 4, the model relies on this parameter



(A) Ablation results for 1 week ahead.

(B) Ablation results for 2 week ahead.

(C) Ablation results 3 week ahead.

(D) Ablation results 4 week ahead.

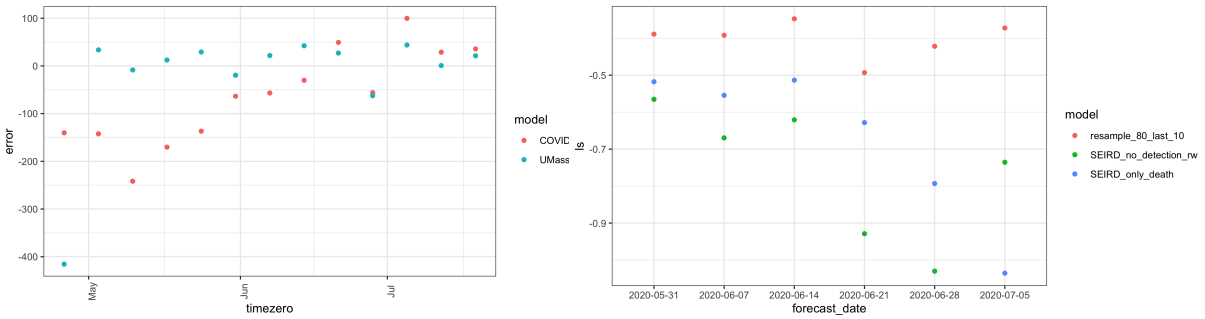
FIGURE 6. Nested model comparison MAE broken down by timezero and target. We can see that MechBayes Full performs better than MechBayes No Detection Random Walk and MechBayes Only death in almost all breakdowns. We can also see that the improvement becomes more pronounced at larger horizons, suggesting that MechBayes Full is a better longer term forecasting model.

heavily, setting the detection probability at around 15% at the start of the pandemic in March, to nearly 80% by August 1st 2020. There is remarkable similarity between the increase across geographical locations as well, suggesting (as testing data reflects) an overall increase in the number of detected cases across the U.S.

- **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 interventions. 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 true changes in transmissibility. Bypassing the epidemiological interpretation that compartmental models provide.
- **Models must evolve in order to be successful in real-time pandemic forecasting**. As Figure 5 shows, MechBayes significantly improved over time relative to the baseline model. Forecasting during a real-time pandemic is subject to a wide variety of data reporting issues. Continuously responding to challenges is key to being able to forecast well.
- **MechBayes is biased high**. We can see from Figure 7 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 COVID-HUB 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**

- **MechBayes is probabilistically well-calibrated compared with the baseline model**. Small WIS values are better in terms of calibration [26]. 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.
- **MechBayes Full outperforms the other versions on probabilistic scoring measures**. As can be seen by Figure 7, MechBayes Full achieves the best log-score **sub for WIS** out the three nested models. This suggested that our extensions to the basic SEIR model not only improve point-forecasts, but also calibration.
- **MechBayes is remarkably accurate**. As we can see from Figure 5 the MAE values, when averaged over geographical unit, are as low as 20 deaths. On average the MAE values are near 50 deaths. Although there is an open question as to what is accurate enough to be actionable by public health officials, MechBayes seems to be able to provide an indicator within a reasonable range of error to make decisions about resource allocation. However, as expected, there is wide variance by geography due to differing population size. If we normalize by population size, then the highest MAE by geographical unit is no more than 55 deaths per 100,000 people, with an average of 20 deaths per 1000,000 people.



(A) Bias of MechBayes and COVID-Baseline as a function of time.

(B) Comparison of nested models using the probabilistic scoring mechanism

FIGURE 7. A) Bias of MechBayes over time. B) Probabilistic Calibration

8. 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 internal and external evaluation, we have demonstrated the success of MechBayes in forecasting. The model is able to improve over a naive baseline model as well as a naive compartmental model. Allowing for time-varying interventions and detection probability 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 COVID-HUB baseline model in both point and probabilistic forecast evaluations. Our nested model comparisons show that extending the basic SEIR compartmental to real-world pandemic challenges improves forecasting accuracy.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] Nicholas G Reich, Matthew Cornell, Evan L Ray, Katie House, and Khoa Le. The zoltar forecast archive: a tool to facilitate standardization and storage of interdisciplinary prediction research. *arXiv preprint arXiv:2006.03922*, 2020.

[26] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *arXiv preprint arXiv:2005.12881*, 2020.

[27] 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.