Abstract

- Covid has affected XX people
- Forecasts useful for public health resource planning, intervention planning (vaccine trials), and disease burden
- Basic Mechanistic Models unable to capture complexities of real world disease epidemics due to
  - Complexities of interventions
  - Issues with testing and reporting
- We propose a novel forecasting algorithm to overcome
  - Under-reporting of cases
  - Time-varying interventions
- Bayesian end to end estimation using both cases and deaths in numpyro

MECHANISTIC BAYESIAN FORECASTS OF COVID19

GRAHAM GIBSON

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 [**?**]. 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 [**?**]. 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. The model is specified by setting the rates of flow between compartments. While these models have been used since their inception in 1918, the COVID pandemic represents a unique opportunity to explore their properties in real-time. 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 [**?**]. 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 [**?**], COVID-19 cases are often dramatically under-reported, with reporting rates for the U.S. estimated at 20-30% [**?**]. In addition, there has also been severe temporal variation in the percent of symptomatic cases [**?**]. We can also see relatively regular weekly reporting cycles, with reporting dropping off significantly on the weekends. Example incident death data is show in 1.

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 as represented by a set of quantiles defined as

$$\mathbb{Q} = .05, .10, ..., .90, .95 \cup .01, .99 \tag{1}$$

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

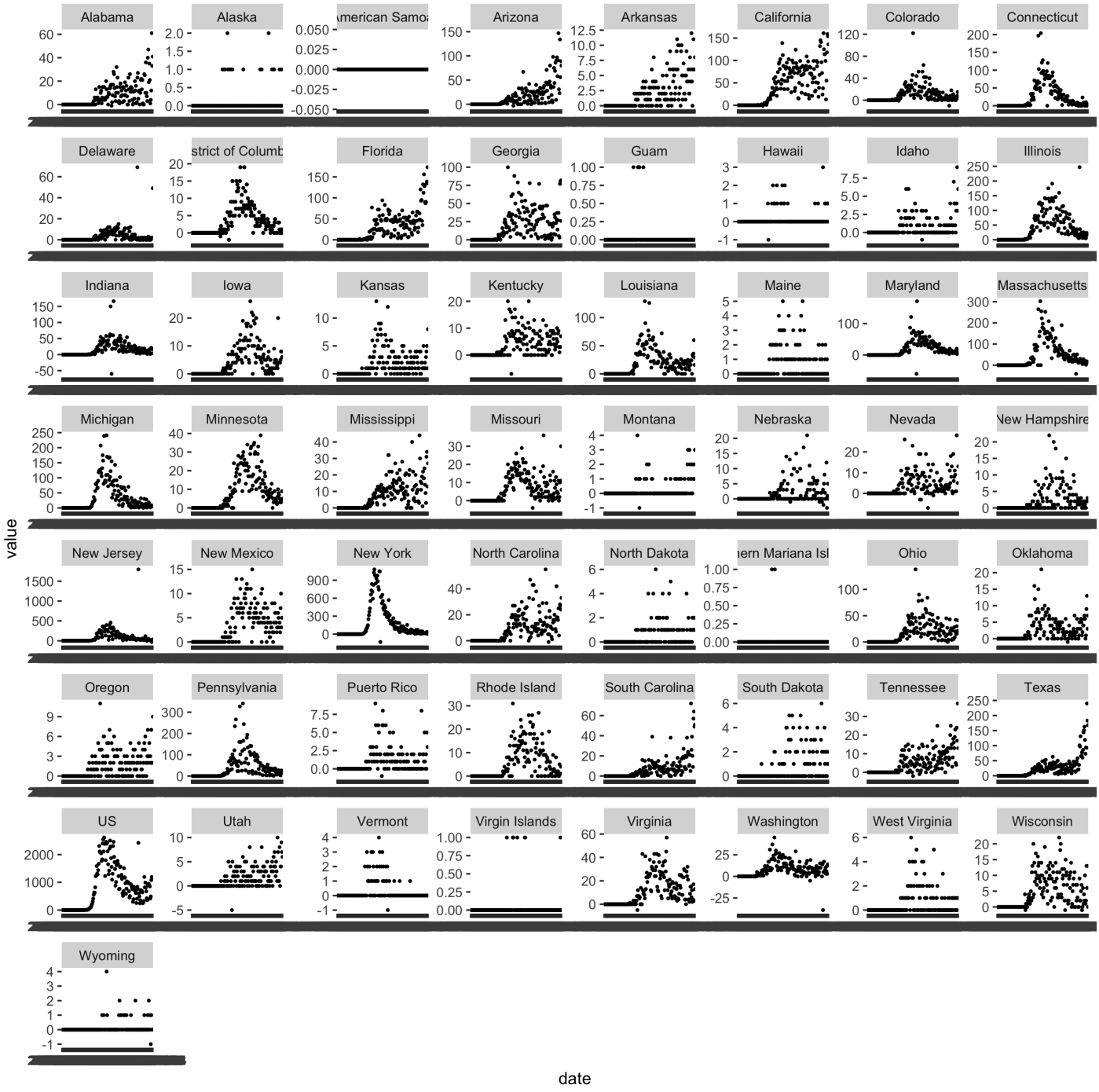


FIGURE 1. Deaths by state for all 50 states and territories. Notice the large variability in incident reporting. There appears to be a weekly cycle where deaths are under-reported on the weekend. We can also see that there are some negative incident deaths, where data are revised to account for deaths that were incorrectly attributed to COVID.

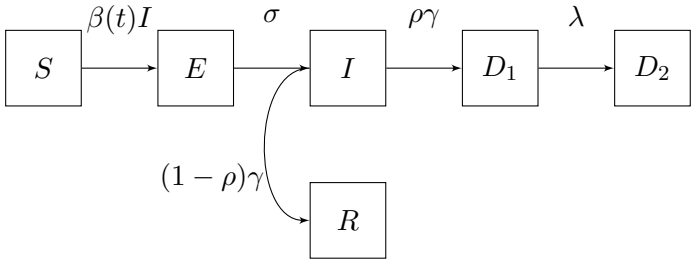


FIGURE 2. Comparmental model parameters

- 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)

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.

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.

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

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.

$$\text{Cases}_t \sim NB(p_{c,t} * I_t, \sigma_c^2) \tag{3}$$

$$\text{Deaths}_t \sim NB(p_d * D_{2_t}, \sigma_d^2) \tag{4}$$

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.

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

This allows us to initialize the process model:

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

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

$$
\begin{aligned}
\sigma &\sim \Gamma(5, 5\hat{d}_E) \\
\gamma &\sim \Gamma(7, 7\hat{d}_I) \\
\beta(0) &\sim \Gamma(1, \hat{d}_I/\hat{R}) \\
\rho &\sim \text{Beta}(10, 90) \\
\lambda &\sim \Gamma(10, 100)
\end{aligned}
$$

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 λ $\mathbb{E}[\lambda] = 0.1$ 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 [**?**]. 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. **I want to say a little more**

3.5. **Fitting.**

- Fully Bayesian HMC on parameters
- Why we choose deterministic compartmental model with only uncertainty on parameters and observations
- Estimation in numpyro very fast

4. Experimental Setup

4.1. **COVID-HUB.**

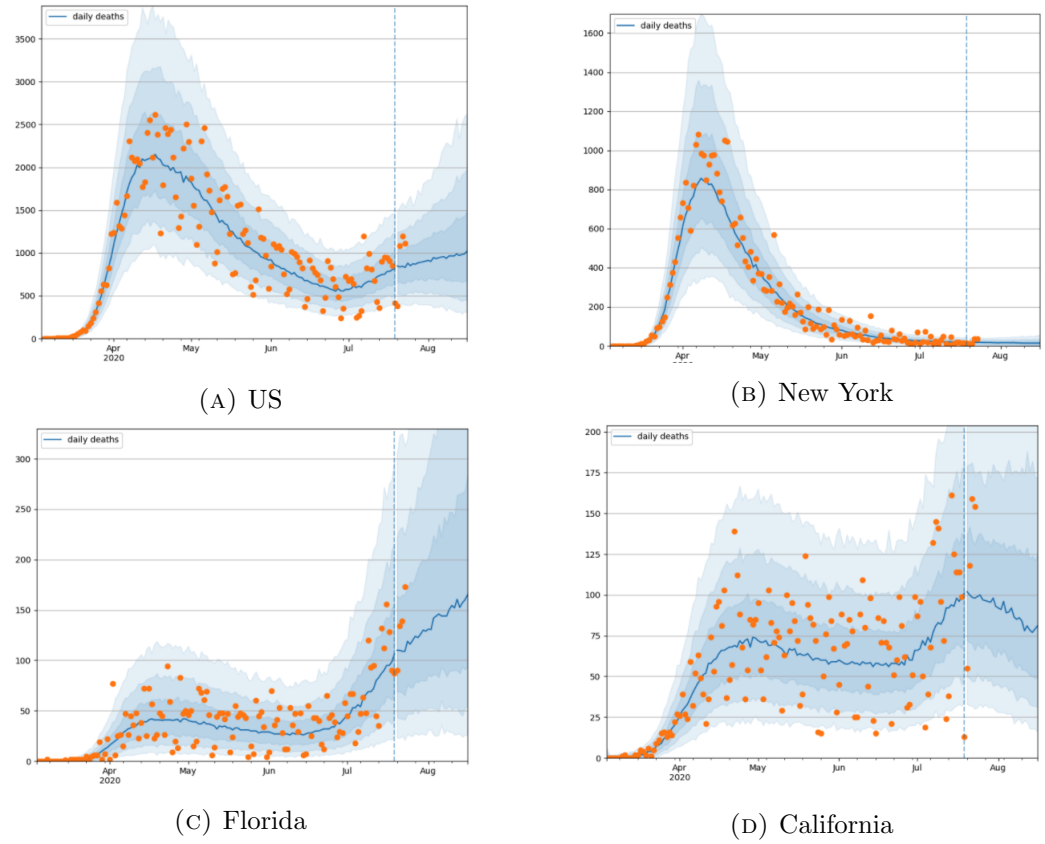
- Real-time forecasting evaluation for 14 weeks starting April 20th 2020.
- Forecasts submitted every Monday using incident data up until Sunday
- Cumulative forecasts generated by aggregation of incident forecasts
- 1-4 week ahead targets generated from 28 day ahead predictions

4.2. **Ablation Test.**

- Evaluate a set of nested models
  - Basic SEIR model with observations only on deaths
  - SEIR with joint observations
  - SEIR with joint observations and random walk detection probability
- We omit the test that involves no random walk on beta since a model must take into account interventions.

5. COVID-Hub Results

- MechBayes is almost always better than baseline model when broken down by region, target, and timezero after June 1st 2020.
- MechBayes has improved over time.
  - First two submissions did not have detection probability random walk
  - Observations were Normally distributed on cumulative deaths
  - Switched to current model
- Naive baseline is hard to beat
- MechBayes is biased high. This comes from uncertainty in the random walk leading to potentially huge growth rates.



(A) US

(B) New York

(C) Florida

(D) California

FIGURE 3. Example fit and forecast for four states.

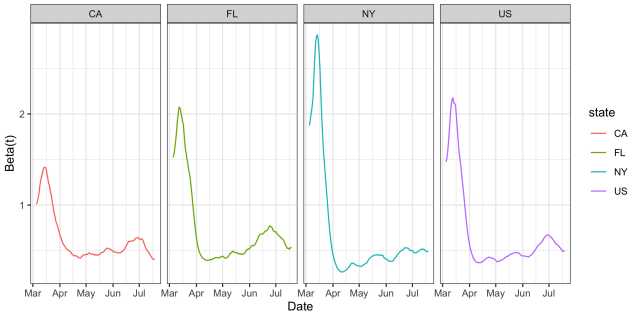


FIGURE 4. Time varying transmissibilty par

6. Ablation Results

- Model ranking is as follows

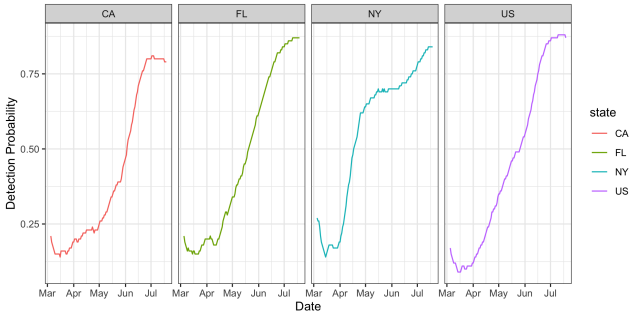


FIGURE 5. Time-varying detection probability.

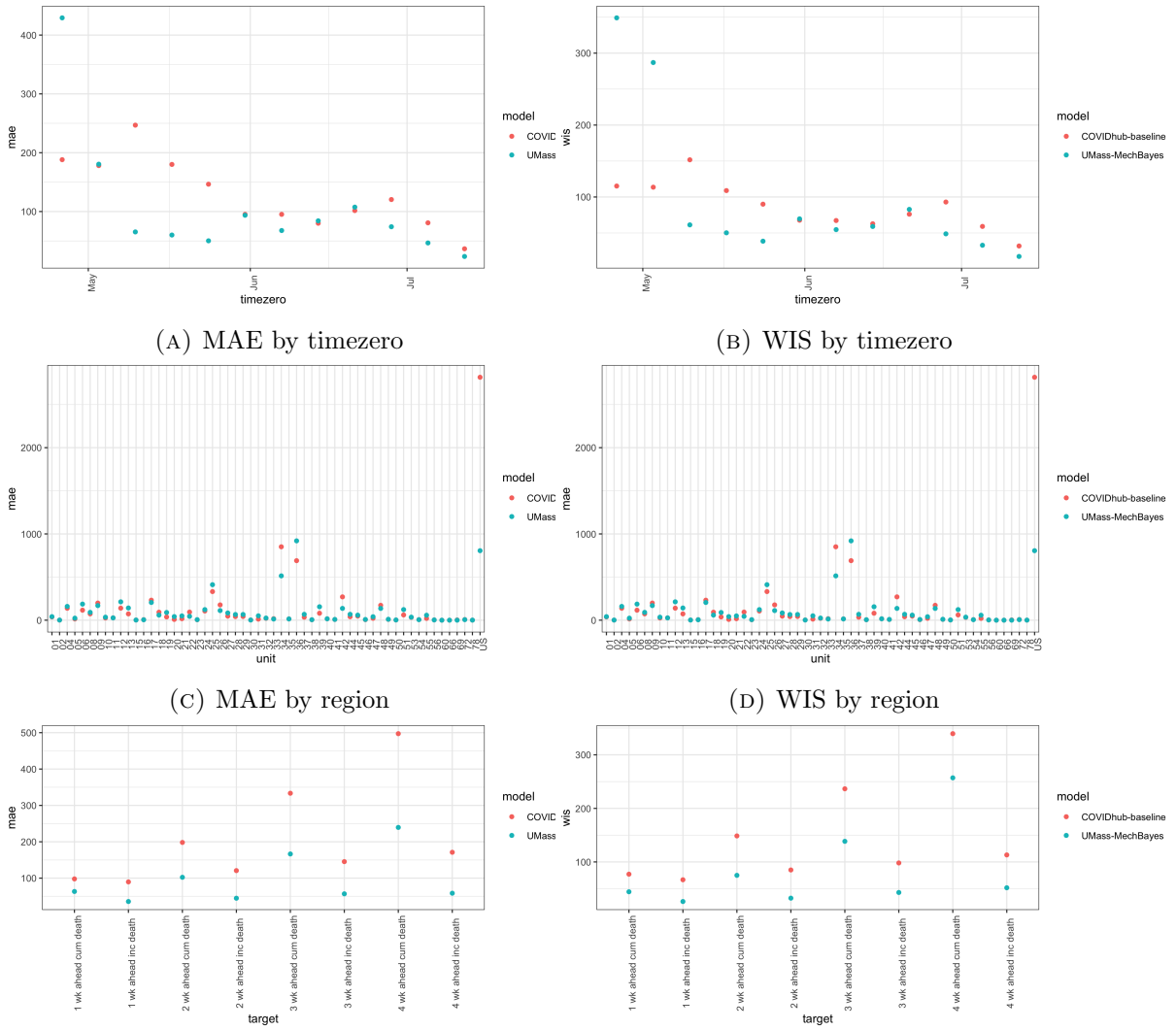
  - MechBayes
  - MechBayes-Case Observations
  - MechBayes - no detection random walk
- This ordering is inu

7. Discussion

- Mech Bayes is a fast fully bayesian compartmental model capable of accounting for real-world modeling challenges during a pandemic.
- Demonstrated success across regions, targets, and timezeros
- Real-time model results show the practice of modeling during an epidemic. Results are improving.
- Ablation studies show the results are grounded in real model improvements using historical validation.
- Talk about how overall MAE obscures the huge geographic variability.

8. Conclusion

- Summarize mech bayes as bayesian compartmental model
- Further work: most of the model is about $\beta(t)$. Better methods for modeling it? Spline Etc.



(A) MAE by timezero

(B) WIS by timezero

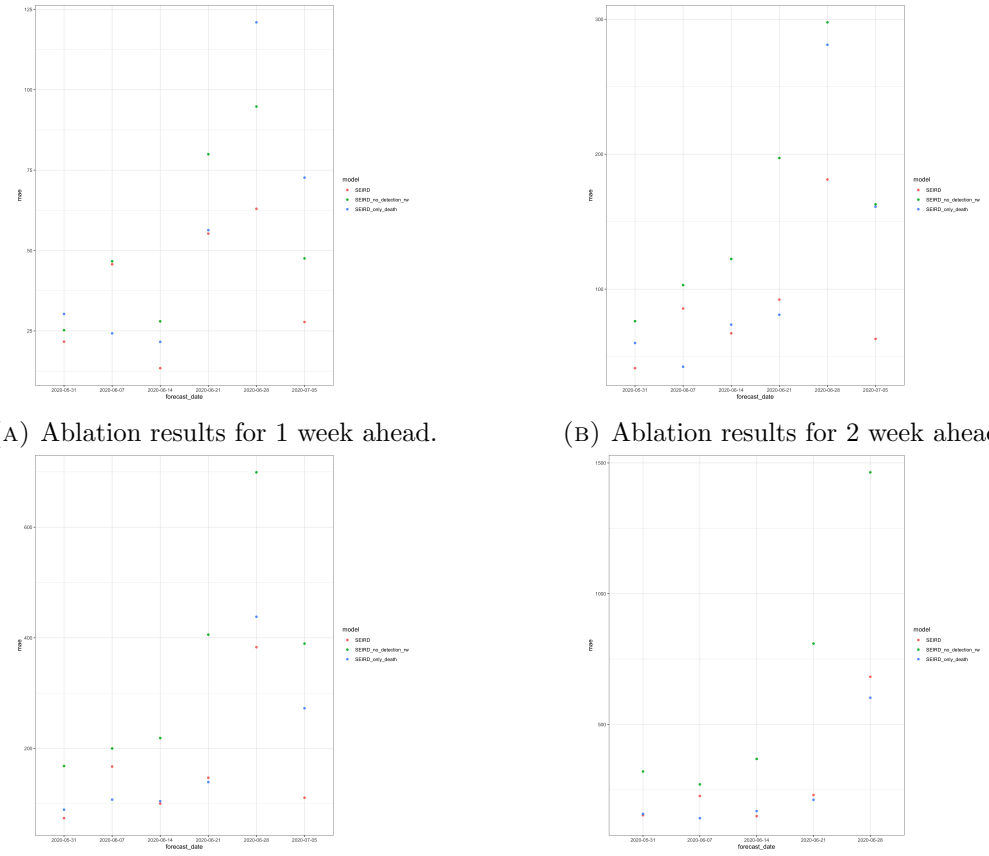
(C) MAE by region

(D) WIS by region

(E) WIS by target

(F) WIS by target

FIGURE 6. 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.



(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 7. 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.