State-specific Projection of COVID-19 Infection in the United States and Evaluation of Three Major Control Measures

Shi Chen[1], Qin Li[1,*], Song Gao[2,*], Yuhao Kang[2], and Xun Shi[3]

[1]Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA
[2]GeoDS Lab, Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA
[3]Department of Geography, Dartmouth College, Hanover, NH 03755, USA
[*]Corresponding Authors: qinli@math.wisc.edu; song.gao@wisc.edu.

ABSTRACT

Most models of the COVID-19 pandemic in the United States do not consider geographic variation and spatial interaction. In this research, we developed a travel-network-based susceptible-exposed-infectious-removed (SEIR) mathematical compartmental model system that characterizes infections by state and incorporates inflows and outflows of interstate travelers. Modeling reveals that curbing interstate travel when the disease is already widespread will make little difference. Meanwhile, increased testing capacity (facilitating early identification of infected people and quick isolation) and strict social-distancing and self-quarantine rules are most effective in abating the outbreak. The modeling has also produced state-specific information. For example, for New York and Michigan, isolation of persons exposed to the virus needs to be imposed within 2 days to prevent a broad outbreak, whereas for other states this period can be 3.6 days. This model could be used to determine resources needed before safely lifting state policies on social distancing.

Introduction

The Coronavirus disease (COVID-19) is an ongoing pandemic that poses a global threat. As of March 26, 2020, more than 520,000 cases of COVID-19 have been reported in over 200 countries and territories, resulting in approximately 23,500 deaths[1–9]. In the United States, the first known positive case was identified in Washington state on January 20, 2020[10]. By March 26, the epidemic had been rapidly spreading across many communities and present in all 50 states, plus the District of Columbia; the total number of confirmed cases in the United States rose to 78,786 with 1,137 deaths.

To combat the spread of COVID-19, the government has taken actions in various dimensions, including banning or discouraging domestic and international travels, announcing stay-at-home orders to curb non-essential interactions for reducing transmission rate, and urging commercial laboratories to increase test capacity. To curb traveling, on January 31, the United States government announced travel restrictions on travelers from China; on February 29, it announced travel ban against Iran and advised travel with caution to Europe[11] ; on March 11, it announced travel restrictions on most of European countries. To reduce human-interactions, on March 13, a national emergency was declared; as of March 28, 39 states had issued either statewide or regionally stay-at-home or shelter-in-place order, requiring residents to stay indoors except for essential activities. To increase test capacities, on February 4, the United States Food and Drug Administration (FDA) approved the United States Centers for Disease Control and Prevention (CDC)'s test, which was later to be proved inconclusive[12]; on February 29, the FDA relaxed its rules for some laboratories, allowing them to start testing before the agency granting its approvals; on March 27, FDA issued an Emergency Use Authorization to a medical device maker, the Abbott Labs, for the use of a coronavirus test that delivers quick testing results[13].

So far, since there is no treatment or vaccine for SARS-COV-2 available, these actions have been taken largely based on classic non-pharmaceutical epidemic controls. Works on evaluating similar measures in other countries, especially China, started to emerge[7,14,15]. For example, the effect of travel restriction on delaying the virus spread in China has been reported[5,16]. However, it is still unclear what control and intervention measures would have actual effect, especially to what extent, on abating the spread of COVID-19 in the United States. As the United States has very different political, administrative, social, pubic health and medical systems, as well as culture from China, this remains to be a critical question to address, especially considering that some measures and policies come with extremely high economic and societal costs.

There have been numerous modeling works projecting or predicting the trend of the COVID-19 pandemic regionally or globally[17,18]. Most of the works apply a global model to the entire study area, either a region, a country, or the entire globe. Rarely the variation of different parts within one area and the interactions among those parts are taken into consideration.

However, a country like the United States features diversity in all aspects. On the one hand, the overall situation of the entire country is a result emerging from local situations and their interactions, and thus, ignoring the local interactions can hardly lead to a high-quality overall model; on the other hand, as all interventions and policies finally have to be adapted to the local situation, a localized modeling will be much more relevant to the real-world practices. Spatially and network-related epidemic models can describe the geographical spread of viral dynamics[7,19–21]. Recent studies have shown the importance of incorporating timely human mobility patterns derived from mobile phone big data and global flight networks into the epidemiology modeling process and in public health studies[5,7,22–30]. Without accurate models that incorporate human mobility patterns and spatial interactions[26,27], it is rather challenging to quantify the sensitivity of parameters, and using the linkage to real practices to make sensible policy suggestions.

Accordingly, the core of the study is twofold. First, to localize the modeling, we developed a travel-network-based susceptible-exposed-infectious-removed (SEIR) mathematical compartmental model system that simultaneously characterizes the spatiotemporal dynamics of infections in 51 areas (50 states and the District of Columbia). Each state or district has its own model, and all models simultaneously take into account inflows and outflows of interstate travelers.

Second, to improve the practical relevance, we chose to use three parameters that can directly correspond to possible practical means to discover, combat, and control the spread of the disease, and quantify their impact on the final output of the model. The three parameters include: 1) the transmission rate b, which corresponds to the local social-distancing enforcement, e.g., the stay-home order; 2) the detection and reporting rate r, which corresponds to the testing capacity; and 3) the travel ratio α_t, which corresponds to the ratio of interstate travel volume compared to that of 2019 during the same period.

The modeling is a dynamic projection process (see the 'methods' section). We employed daily and state-specific historical data to incrementally calibrate the model, and then used the calibrated model to predict future scenarios under different non-pharmaceutical control and intervention measures. During this process, we ran data assimilation methods to identify parameter values that optimally fit the current situation (see more details in the methods and supplementary material). To project into the future, we set different values for the parameters to create different control and intervention scenarios, and then ran the simulation to see their impact on the model results. The final output of the model is the total number of confirmed cases in a state on a particular day. The current strategy in the United States is to isolate people who have the symptoms of COVID-19. An ideal scenario is to have an 100% reporting rate, i.e., every infected case gets confirmed and thus isolated quickly. Another ideal setting is to have everyone who was in contact with the infected gets identified and isolated quickly as well. Our model incorporated these considerations and examined such direct isolation of the exposed compartment in detail. We particularly investigated the impact of quickness of such actions through mathematical modeling and scenario analysis.

A notable result from our modeling is that the impact of interstate travel restriction on the model output is modest. This can be explained by that when the disease has already widespread in all states, the relatively small number of cases in the travelers will cause little difference to the local situation, compared with the effects of local social-distancing and isolation rules and the increase of testing capacity.

Results

Fig. 1 shows the effect on spatiotemporal dynamics of infectious population across states by setting the coefficients at different configurations. An interactive map-based scenario simulation web dashboard is also available at `https://geods.geography.wisc.edu/covid19/us_model`. We set $r = 1 - \alpha_r(1 - r_0)$ and $b = \alpha_b b_0$, where r_0 and b_0 are the report and transmission rate as of March 20, 2020 using data assimilation fitting result. By decreasing α_r from 1 to 0, we increase the report rate from the original r_0 to 1, and by decreasing α_b we decrease the transmission rate. Most states, except a few such as NY, MI, and CA, see drastic improvement when the transmission rate is decreased and the testing(reporting) rate is increased, but the reduction of interstate traffic alone is not as effective. Our modelling reveals that once the epidemic in an area has reached a certain stage, the difference that can be caused to the local situation by the relatively small number of imported cases due to the interstate travel is insignificant. According to our modeling, all states in the United States have reached that stage. Therefore, as long as those travelers follow the social-distancing rules and the local government provides sufficient testing capacity, there is no apparent urge to curb interstate travel. This is in line with the finding in[16,28], in which the authors projected the pick up of the spreading in other parts of China outside of Wuhan with about 3 days delay, and in the world outside China within a 2-3 weeks of delay, assuming no further screening is in place. Different from China where the city of Wuhan is clearly the epicenter of the COVID-19 outbreak and the travel ban quickly gets the rest of China under control, most of the states in the United States have already had signs of community spread by March 20, 2020[31], and banning other states will hardly make much difference to the local situation. In addition, Fig. 2 shows the corresponding prediction time series of infectious population in top 15 states under two scenarios (see also Fig. S14): (A) the reported rate and the transmission rate remained unchanged as of March 20, 2020, with $\alpha_r = \alpha_b = 1$, in which most states will continue their exponential growth before reaching their peak; (B) with $\alpha_r = \alpha_b = 0.1$, that is, when the transmission rate b is much smaller and the reported rate r is much higher (closer to 1), we can "flatten the curve" on the virus (i.e., reducing the spread of the virus).

We further investigate the effect of increased testing capacity and report rate. As shown in Figure 3a, most states see drastic improvement when the report rate increases. All states, by April 29, see monotonically exponential reduction of infections. The impact is strong in states such as MA, AZ, FL, and OR, but relatively weak in states such as NY, MI and IL. In Figure 3b, we study the effect of α_r and α_b on the basic reproduction rate R_e in NY (see other states in Fig. S15). It can be seen that merely raising the report rate cannot fully make $R_e < 1$. To mitigate the spread of COVID-19 in these states, a proactive approach needs to be taken, and quick detection and isolation of the exposed population need to be in place instead of being delayed until the onset of the symptoms. This measure can prevent the exposed population from potentially infecting other susceptible people. In Figure 3c, we plot the increase of infections in terms of D_q (i.e., the temporal lag in putting a person into quarantine) for the states that are sensitive to change of D_q, including NY, NJ, IL, GA, MI, CO, WI, LA, TX, PA, MA, and TN. The longer one waits to inform and isolate the exposed population, the more infected people one observes. For example, there is a sharp transition for NY and MI. If the average detection and isolation time is more than 2 days, the total number of infections will significantly increase.

The results again showed the importance of sufficient testing and strong transmission-intervention measures such as social distancing and self-quarantine policy[32]. These policies can help quickly identify the source of infection and isolate them before they infect the remaining population. This measure presumably comes with a lower economical cost.

We finally investigate the stability of our statements on the parameters chosen in the model. There are a number of parameters in the model that are determined according to medical studies and thus necessarily contain ambiguity. One parameter, γ, is especially hard to be set at a particular value due to the lack of medical evidence. This parameter reflects the level of infectiousness of the “exposed" compartment, a population that is presymptomatic. Recent studies indicate that presymptomatic patients seem to be more infectious than patients who have symptoms on site[33]. We therefore run our model with different values of γ to identify the significance of this particular parameter. Our numerical result suggests that within a moderate range of γ, our conclusions still stand true. In particular, as shown in Figure 4, by setting the “exposed" compartment being more infectious than the “infected" compartment, the numerical solution shows the same trend. We still observe that, with a higher report rate, the number of non-infected population exponentially increases (i.e., less people would get infected), and when a proactive approach is taken, meaning that the “exposed" compartment gets quickly separated from the rest of the population, the non-infected population drastically increases as D_q, the delay of the separation time, gets shortened. This means that the dependence of our conclusion on the parameter γ is stable, and the above statements are consistent.

We should emphasize that in our simulation, we do not differentiate patients with severe or mild symptoms. A more dedicated numerical experiment that separates the two categories could potentially give more detailed information. For example, in another agent-based modeling study[34], researchers consider patients with mild to severe symptoms to evaluate the impacts of the timing of social distancing and adherence level on COVID-19 confirmed cases.

Discussion and conclusion

Modeling and analyzing the spread of COVID-19, and assessing the effect of various policies could be instrumental to national and international agencies for health response planning[5,8,15–17,32]. We show that the effect of interstate travel reduction is at most modest in the United States when the outbreak has already widespread in all states. On the other hand, we need to impose strong transmission-reduction intervention and increased testing capacity and report rate to contain the spread of virus. The result is based on mathematical and statistical analyses of transmission control measures and in agreement with previous findings[2,3,5,14–16], suggesting that the effect of travel ban at a later stage of the outbreak is rather modest. This is also in line with the fact that the outbreaks still occurred in Europe even upon the strong travel ban on the earlier epicenter of Wuhan and its surrounding cities in China. We also quantitatively show that the transmission-reduction intervention such as policies on the social-distancing and shelter-in-place rules, and the increase of testing rate, which facilitates immediate isolation upon exposure, will significantly reduce the total infected population. Such effect is mostly visible for the states of NY, NJ, MI, and IL. Particularly, our modeling results show that for states such as NY and MI, to achieve an optimal infection reduction, a more proactive approach needs to be taken to quickly identify the exposed population and isolate them within two days of exposure in order to ensure the infection reduction. The result is in agreement with previous findings[7,8].

We do need to emphasize that the model itself does not distinguish different ways of traveling across states. Indeed, if the interstate travel is conducted mostly through transiting through busy airports and train stations, and the social-distancing policy is not strictly imposed, then the high population density at these places will bring up the transmission rate b locally in space and time, leading to a higher infection rate. This is a severe consequence, but it should not be counted as the direct result of relaxing travel restrictions.

Moving forward, we estimate that the decline in travel has a modest effect on the mitigation of the pandemic. We need a stronger transmission-reduction intervention and increased detection and report rate in place to prevent the further spread of the virus. The results could potentially be used to design an optimal containment scheme for mitigating and controlling the spread of COVID-19 in the United States.

Methods

The mathematical model that simulates the spatiotemporal dynamics of state-level infections in the United States is a modified travel-network-based SEIR compartmental model in epidemiology by taking into account the variation of the 51 administrative units and their interactions[14,35–37]. It consists of 51 ordinary differential equation (ODE) systems, with each one characterizing the evolution of susceptible (S), exposed (E), reported (I), unreported (U) and removed (R) cases per state (Fig. S1 and see more details in the supplementary material). The 51 ODE systems are then coupled through the state-to-state travel network flows (see Fig. S2) that were extracted from the aggregated SafeGraph mobility data and weighted by α_t[38,39]. Unlike most other models, we also incorporate the potential asymptomatic transmission. This makes the derivation of the basic reproduction number R_0 different. Besides, each ODE system also includes two unknown parameters: the transmission rate (b) and the report rate for each state (r). The unknown parameters are inferred based on the total number of confirmed cases in each state for the period of March 1-March 20, 2020. The source of infection case data is the Center For Systems Science and Engineering at the Johns Hopkins University[9].

The parameters and model specification are defined as follows:

$$\begin{cases} \frac{\mathrm{d}S_i}{\mathrm{d}t} = -\frac{b_i S_i (U_i + \gamma E_i)}{P_i} + \sum_{j \neq i} \alpha_t n_{ij} \frac{S_j}{P_j} - \sum_{j \neq i} \alpha_t n_{ji} \frac{S_i}{P_i} \\ \frac{\mathrm{d}E_i}{\mathrm{d}t} = \frac{b_i S_i (U_i + \gamma E_i)}{P_i} - \frac{E_i}{D_e} + \sum_{j \neq i} \alpha_t n_{ij} \frac{E_j}{P_j} - \sum_{j \neq i} \alpha_t n_{ji} \frac{E_i}{P_i} \\ \frac{\mathrm{d}I_i}{\mathrm{d}t} = r_i \frac{E_i}{D_e} - c_I \frac{I_i}{D_c} - (1 - c_I) \frac{I_i}{D_l} \\ \frac{\mathrm{d}U_i}{\mathrm{d}t} = (1 - r_i) \frac{E_i}{D_e} - c_U \frac{U_i}{D_c} - (1 - c_U) \frac{U_i}{D_l} + \sum_{j \neq i} \alpha_t n_{ij} \frac{U_j}{P_j} - \sum_{j \neq i} \alpha_t n_{ji} \frac{U_i}{P_i} \\ \frac{\mathrm{d}R_i}{\mathrm{d}t} = c_I \frac{I_i}{D_c} + (1 - c_I) \frac{I_i}{D_l} + c_U \frac{U_i}{D_c} + (1 - c_U) \frac{U_i}{D_l} \end{cases} . \tag{1}$$

The ODE system is equipped with the following initial data ($t = 0$ standing for March 1, 2020):

$$S_i(0) = N_i - E_{i0} - U_{i0} - I_{i0}, \quad E_i(0) = E_{i0}, \quad I_i(0) = I_{i0}, \quad U_i(0) = U_{i0}, \quad R_i(0) = 0. \tag{2}$$

In the equation, the unit for t is one day. $N_i(t)$ is the total population of state i at time t, and $P_i = S_i + E_i + U_i$ is the free population. n_{ij} is the number of inflow from state j to state i. b_i and r_i are the transmission rate and reporting rate of state i. c_I (c_U, resp.) is the proportion of positive cases that show critical condition for I (unreported cases U, resp.). D_e is the latent period. D_c and D_l are the infectious periods of critical cases and mild cases. α_t is a parameter to tune the traffic flow.

We emphasize two main differences in modeling compared with existing literature. In[7], the authors study the inter-city traffic and its impact on the spreading of COVID-19 in China. The situation in China and that in the US are very different. In China, the epicenter is clear: the city of Wuhan, Hubei province, and the outbreak starts mid-January, 2020. The COVID-19 outbreak in the US, however, is multi-sourced. The consequence is that in the model in[7], the initial condition for cities excepts Wuhan is clear: the latent, the reported and the unreported cases are all zero. In this model, however, the initial conditions E_{i0} are unclear for all states; Another big difference is, according to clinical findings, the latent cases also have the potential of transmitting the virus, and thus we add the interaction of E_i with S_i into the increment of E_i[7,40,41].

The unknown parameters and state variables in the equation set are

* b_i: the transmission rate with non-informative prior range $[1, 1.5]$;
* r_i: the report rate with non-informative prior range $[0.1, 0.3]$;
* E_{i0}: the data for the latent population with non-informative prior range $[0, 500]$.
* U_{i0}: the initial data for the unreported population with non-informative prior range $[0, 200]$.
* S_{i0}: the initial data for the susceptible population defined by $N_i - E_{i0} - I_{i0} - A_{i0}$.

Other parameters are:

γ: the transmission ratio between unreported and latent. In the simulation we set it to be 0.5;

D_c: the average duration of infection for critical cases. We assume $D_c = 2.3$ days[42].

D_e: the average latent period. According to[43], $D_e = 5.2$ days.

D_l: the average duration of infection for mild cases. We assume $D_l = 6$ days.

α_t: the ratio of interstate travel volume compared to that of 2019 during the same period. The travel flow information n_{ij} was extracted from the SafeGraph mobility data, and we set $\alpha_t = 0.5$ to represent the travel reduction situation observed in the year of 2020.

c_I: proportion of critical cases among all reported cases. We choose $c_I = 0.1$

c_U: proportion of critical cases among all unreported cases. We assume $c_A = 0.2$.

There is an essential assumption made in the model: the homogeneity in the population. It means that the traffic flow is a good representation of the total population without considering their demographic and socioeconomic characteristics. The susceptible, exposed, and unreported move in and out of states at the same rate. This explains the $\frac{S_i}{P_i}$, $\frac{E_i}{P_i}$ and $\frac{U_i}{P_i}$ terms in the $S_i/E_i/U_i$ equation.

The effective reproductive number R_e could be computed as

$$R_e = \frac{b}{E+U}\left[\gamma D_e E + \frac{D_c D_l U}{c_U D_l + (1-c_U) D_c}\right]. \tag{3}$$

R_e depends on time due to the time dependence of E and U.

The COVID-19 transmission dynamics (the ODE system) was simulated using the Forward Euler method, with each day discretized into 24 smaller time periods to ensure the numerical stability (see Fig. S3). The parameter fitting was conducted under the Bayesian formulation that combines the effect of the underlying dynamics governed by the ODE system, serving as the prior knowledge, and the collected data, appearing in the likelihood function, to generate the posterior distribution that characterized the behavior of the state variables, including S, E, I, U, R, as well as the two unknown parameters, b and r. For this classical data assimilation problem, we employed the Ensemble Kalman Filter method that was derived from the Kalman filter and tailored to deal with problems with high-dimensional state variables[44,45]. The method proves to be effective when the measuring operator is linear and the underlying dynamics is Gaussian-like. It has been applied to a vast of problems that do not strictly satisfy the Gaussianity requirement. To apply this method, we generated 2000 samples according to the prior distribution, and evolve the samples through the dynamics of the ODE system. The samples were then rectified at the end of each day, using the announced number of confirmed cases, for tuning the two unknown parameters b and r.

At the beginning of the simulation, March 1, only a few states had non-zero confirmed cases. The true numbers of exposed people and unreported cases on that day, however, are unknown. These two numbers are also the state variables that need to be inferred to using the collected infection data. On March 1, we put a non-informative prior with range $[0, 500]$ and $[0, 200]$ over the exposed latent population and unreported infectious population in each state, respectively. Fig. S4-S13 show the data assimilation results for different states including the number of people in different compartmental groups and their temporal changes with 95% credible intervals. The average reporting rate r over all states is 0.2266 at the end of March 20 through the data assimilation method.

For forecasting (in supplementary material), we performed scenario studies of two types. First, we ran the mathematical model by applying the initial data obtained as of March 20 into the future for the next 40 days, but with different configurations of (b, r, α_t). The simulation results out of this setting were then compared with those from the setting that the three parameters remained unchanged for each state. To quantify and visualize the difference, we compared the increase of the percentage of the non-affected population when the measures of stay-at-home, increasing test rate, and travel bans were enacted.

The second scenario was about a more ideal situation: every confirmed case would get isolated immediately, as well as those who had been exposed to those confirmed cases, no matter if those who had been exposed had started to show symptoms or not. We built a new mathematical model that incorporated such isolations to study the effect of them. A new quarantined compartment (Q) was introduced into the model. Through the simulation, we examined the correlation between the average action-taking time (i.e., temporal lag in putting a person into quarantine denoted by D_q) and the increase of non-infected population. In both scenario studies, the simulation was run with the Forward Euler ODE solver, during which each day was divided into 24 intervals to achieve a numerical stability.

As a SEIR-type epidemic model, this model describes the dynamics of different compartments of the population, and assumes homogeneity within each compartment. However, we should note that this assumption may not be valid in real-world scenarios with heterogeneous populations and infections. Indeed, when an individual contracts the disease, the status could be either mild or severe. In our model, this is absorbed by the report rate r_i but is not explicitly differentiated in the model. A more sophisticated model should have the heterogeneities included, but that would pose a significant higher computational demand and more detailed empirical or clinical data support. We leave that to future research efforts.

Data and code availability

The epidemiological data were retrieved from an open source project: Novel Coronavirus (COVID-19) Cases, developed by the Center For Systems Science and Engineering at the Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data). In addition, we collected millions of points of interest (POIs) with their foot-traffic and anonymous mobile phone users' travel patterns in the United States from SafeGraph. The data for academic research can be requested at https://www.safegraph.com. The code used for modeling and analysis in this paper is available in the GitHub repository: https://github.com/GeoDS/Travel-Network-SEIR/.

References

1. Drake, J. M., Chew, S. K. & Ma, S. Societal learning in epidemics: intervention effectiveness during the 2003 SARS outbreak in singapore. *PloS One* **1** (2006).
2. Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* **395**, 689–697 (2020).
3. Du, Z. *et al.* Risk for transportation of 2019 novel coronavirus disease from Wuhan to other cities in China. *Emerg. infectious diseases* **26** (2020).
4. Tian, H. *et al.* Early evaluation of transmission control measures in response to the 2019 novel coronavirus outbreak in China. *medRxiv* (2020).
5. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).
6. Lipsitch, M. *et al.* Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).
7. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science* (2020).
8. Maier, B. F. & Brockmann, D. Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China. *arXiv preprint arXiv:2002.07572* (2020).
9. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infect. Dis.* .
10. Holshue, M. L. *et al.* First case of 2019 novel coronavirus in the United States. *New Engl. J. Medicine* (2020).
11. Vox news, available at https://www.vox.com/policy-and-politics/2020/2/29/21159273/coronavirus-death-trump-health-officials-travel-ban-iran.
12. New Yorks Times Report, available at https://www.nytimes.com/2020/02/12/health/coronavirus-test-kits-cdc.html.
13. USA Today Report, available at https://www.usatoday.com/story/news/health/2020/03/28/coronavirus-fda-authorizes-abbott-labs-fast-portable-covid-test/2932766001/.
14. Lai, S. *et al.* Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak: an observational and modelling study. *medRxiv* (2020).
15. Ferretti, L. *et al.* Quantifying dynamics of SARS-CoV-2 transmission suggests that epidemic control and avoidance is feasible through instantaneous digital contact tracing. *medRxiv* (2020).
16. Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science* (2020).
17. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infect. Dis.* (2020).
18. Hellewell, J. *et al.* Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Glob. Heal.* (2020).
19. Mollison, D. Spatial contact models for ecological and epidemic spread. *J. Royal Stat. Soc. Ser. B (Methodological)* **39**, 283–313 (1977).
20. Lloyd, A. L. & May, R. M. Spatial heterogeneity in epidemic models. *J. theoretical biology* **179**, 1–11 (1996).

21. Tuckwell, H. C., Toubiana, L. & Vibert, J.-F. Spatial epidemic network models with viral dynamics. *Phys. Rev. E* **57**, 2163 (1998).

22. Meloni, S. *et al.* Modeling human mobility responses to the large-scale spreading of infectious diseases. *Sci. Reports* **1**, 62 (2011).

23. Richardson, D. B. *et al.* Spatial turn in health research. *Science* **339**, 1390–1392 (2013).

24. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).

25. Lai, S. *et al.* Assessing spread risk of wuhan novel coronavirus within and beyond China, January-April 2020: a travel network-based modelling study. *medRxiv* 1–20 (2020).

26. Zhu, X. *et al.* Spatially explicit modeling of 2019-nCoV epidemic trend based on mobile phone data in Mainland China. *medRxiv* (2020).

27. Buckee, C. O. *et al.* Aggregated mobility data could help fight COVID-19. *Sci. (New York, NY)* (2020).

28. Kraemer, M. U. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* (2020).

29. Zhou, C. *et al.* COVID-19: Challenges to GIS with big data. *Geogr. Sustain.* (2020).

30. Grasselli, G., Pesenti, A. & Cecconi, M. Critical care utilization for the COVID-19 outbreak in lombardy, italy: early experience and forecast during an emergency response. *JAMA* (2020).

31. USA Today Report, available at https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

32. Wang, J., Tang, K., Feng, K. & Lv, W. When is the COVID-19 pandemic over? evidence from the stay-at-home policy execution in 106 Chinese cities. *Available at SSRN: https://ssrn.com/abstract=3561491* **106** (2020).

33. Arons, M. M. *et al.* Presymptomatic sars-cov-2 infections and transmission in a skilled nursing facility. *New Engl. journal medicine* (2020).

34. Alagoz, O., Sethi, A., Patterson, B., Churpek, M. & Safdar, N. Impact of timing of and adherence to social distancing measures on covid-19 burden in the us: A simulation modeling approach. *medRxiv* (2020).

35. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. royal society london. Ser. A, Containing papers a mathematical physical character* **115**, 700–721 (1927).

36. Hethcote, H. W. The mathematics of infectious diseases. *SIAM review* **42**, 599–653 (2000).

37. Brauer, F. Compartmental models in epidemiology. In *Mathematical epidemiology*, 19–79 (Springer, 2008).

38. Prestby, T., App, J., Kang, Y. & Gao, S. Understanding neighborhood isolation through spatial interaction network analysis using location big data. *Environ. Plan. A: Econ. Space* 0308518X19891911.

39. Liang, Y., Gao, S., Cai, Y., Foutz, N. Z. & Wu, L. Calibrating the dynamic huff model for business analysis using location big data. *Transactions GIS* (2020).

40. Leung, N. H. *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Medicine* (2020).

41. CNN Report, Infected people without symptoms might be driving the spread of coronavirus more than we realized, available at https://www.cnn.com/2020/03/14/health/coronavirus-asymptomatic-spread/index.html.

42. Guan, W.-j. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *New Engl. J. Medicine* (2020).

43. Wang, C. *et al.* Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in Wuhan, China. *medRxiv* (2020).

44. Evensen, G. The ensemble kalman filter for combined state and parameter estimation. *IEEE Control. Syst. Mag.* **29**, 83–104 (2009).

45. Reich, S. & Cotter, C. *Probabilistic forecasting and Bayesian data assimilation* (Cambridge University Press, 2015).

Acknowledgements

We would like to thank the SafeGraph Inc. for providing the anonymous and aggregated human mobility and place visit data. We would also like to thank all individuals and organizations for collecting and updating the COVID-19 epidemiological data and reports. **Funding:** S.G. and Q.L. acknowledge the funding support provided by the National Science Foundation (Award No. BCS-2027375). Q.L. and S.C. acknowledge the Data Science Initiative of UW-Madison. X.S. acknowledges the Scholarly Innovation and Advancement Awards of Dartmouth College. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Author contributions statement

Research design and conceptualization: Q.L., S.C., S.G.; Data collection and processing: S.C., S.G., Y.H.K.; Mathematical model implementation: Q.L., S.C.; Result analysis: Q.L., S.G., X.S.; Visualization: S.C., S.G., Y.H.K.; Project administration: Q.L. S.G., X.S.; Writing: all authors.

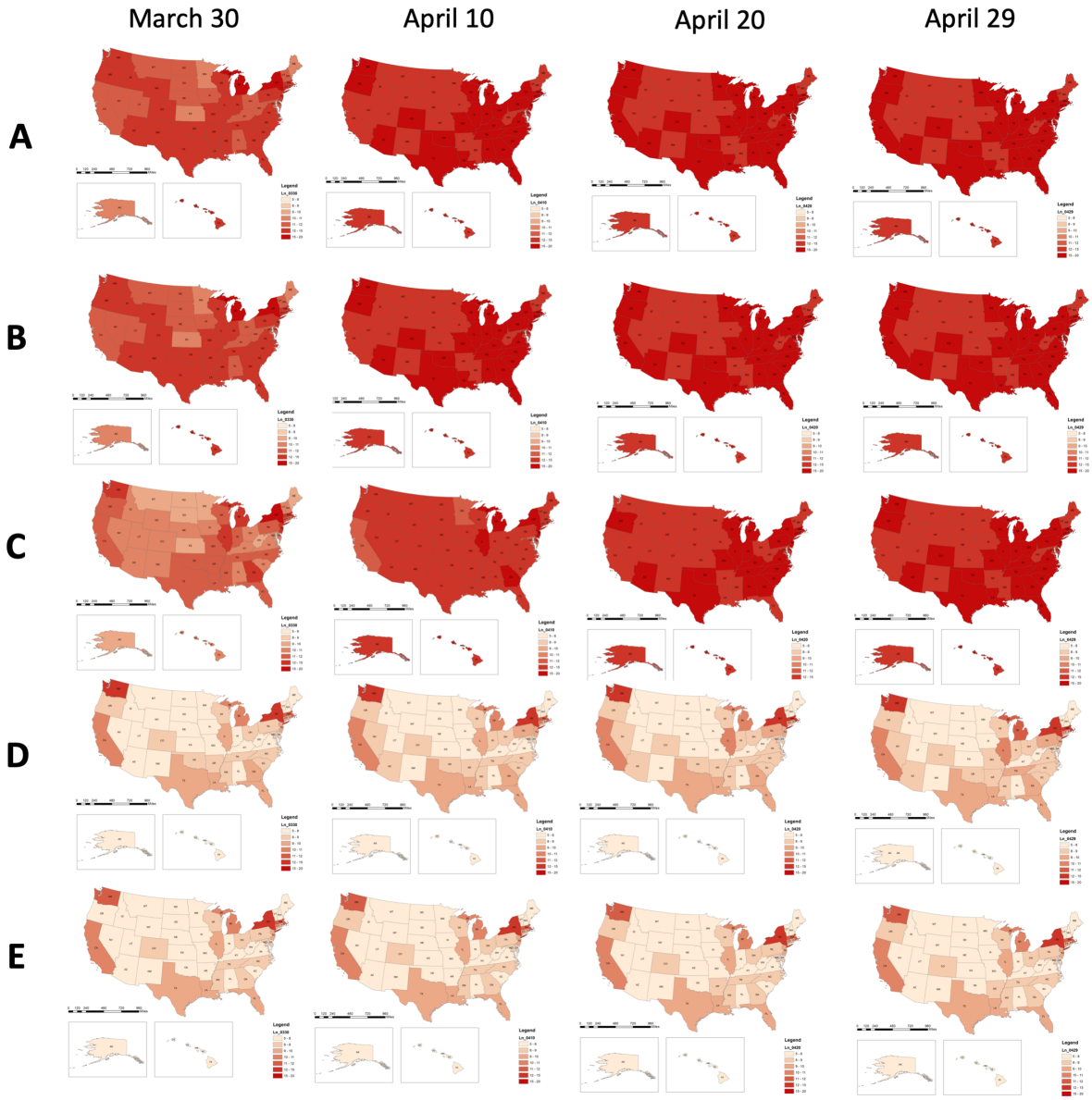


Figure 1. The spatiotemporal distribution of predicted infected population (in natural logarithm scale) across all states under different simulation scenarios: (A) $\alpha_r = 1$ and $\alpha_b = 1$, i.e., all parameters took the values of the initial configuration, obtained through data assimilation method using the numbers of confirmed cases during March 1 – March 20, 2020; (B) the travel flow was reduced to $\alpha_t = 0.05$, while other parameters values remained unchanged; (C) $\alpha_r = 0.1$ and $\alpha_b = 1$; (D) $\alpha_r = 1$ and $\alpha_b = 0.1$; (E) $\alpha_r = 0.1$, $\alpha_b = 0.1$. In the simulations, the transmission rate was set to be $b = \alpha_b b_0$ and the reporting rate was set to be $r = 1 - \alpha_r(1 - r_0)$. Where r_0 and b_0 were the reporting rate and the transmission rate on March 20, 2020, which are inferred from the data assimilation step.

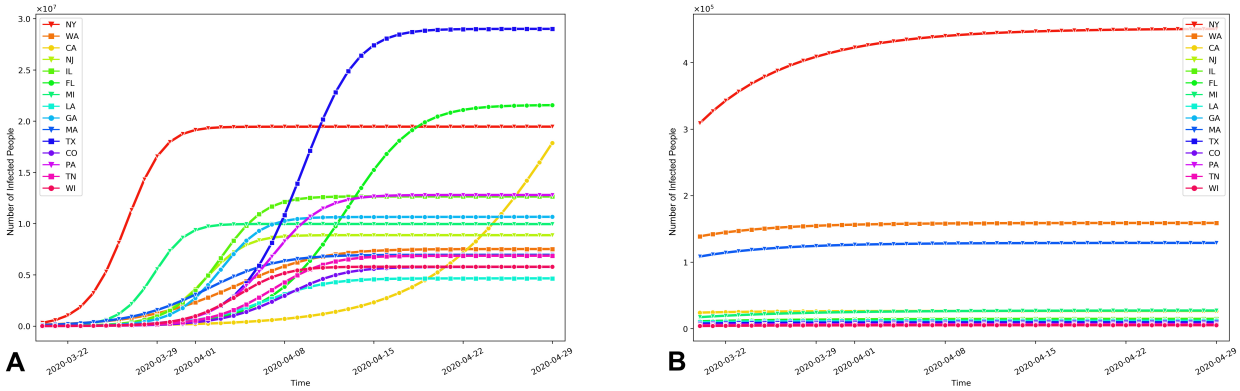


Figure 2. The prediction time series of the total infected population in the 15 most affected states under two scenarios: (A) $\alpha_r = \alpha_b = 1$, i.e., both the reported rate and the transmission rate remained unchanged; (B) $\alpha_r = \alpha_b = 0.1$, i.e., the transmission rate b was smaller and the reported rate r was larger (closer to 1) as $r = 1 - \alpha_r(1 - r_0)$.

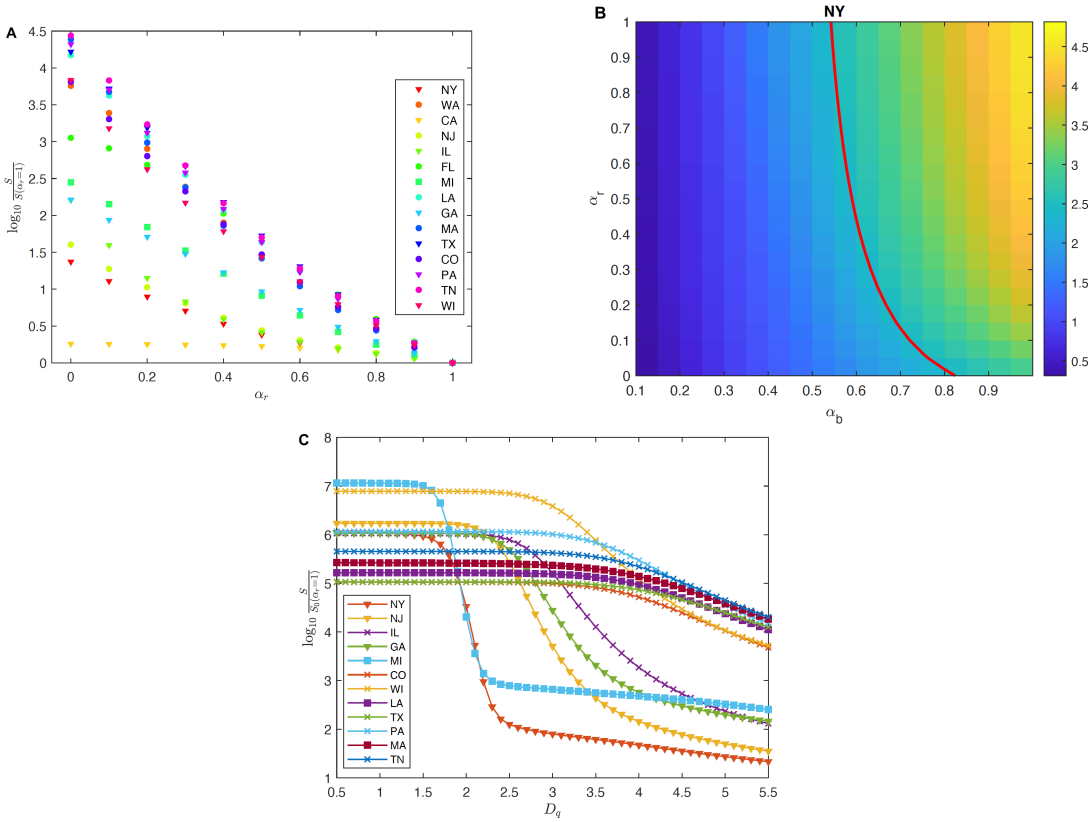


Figure 3. (A) Susceptible population (S) on April 29, 2020 as a function of α_r. $S(\alpha_r = 1)$ is the susceptible population on April 29 computed with the report rate set as the original report rate inferred from the data assimilation step. In all states, S increases as α_r decreases, meaning that more people stay unaffected when a higher report is enacted. (B) R_e, the basic reproduction number, on April 29 for different α_b and α_r in NY. The red line is the level set $R_e = 1$. It can be seen that increasing the reported rate helps diminish the reproductive number, but cannot reduce R_e under 1 if the original transmission rate b_0 is applied; (C) Susceptible population on April 29 for different D_q. $S(\alpha_r = 1)$ is the same as in (A). S significantly depends on the period from expose to quarantine.

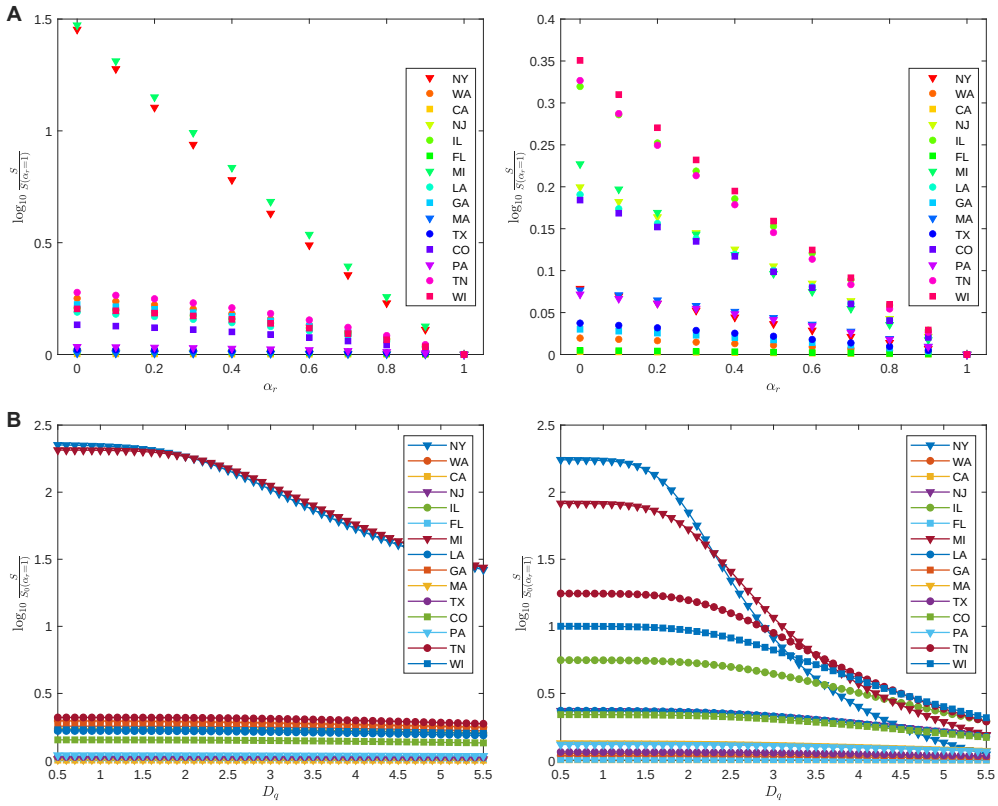


Figure 4. (A) Susceptible population (S) on April 4, 2020 as a function of α_r. The panels on the left and on the right are results from $\gamma = 0.5$ and $\gamma = 1.5$, respectively. For both γ, S increases as α_r decreases, meaning that more people stay unaffected when a higher report is enacted. (B) Susceptible population on April 4, 2020 for different D_q and different values of γ. For those states whose susceptible population is much smaller than their total population due to a high infection rate (such as in NY), S significantly depends on D_q for both $\gamma < 1$ and $\gamma > 1$.

Supporting Material for "State-specific Projection of COVID-19 Infection in the United States and Evaluation of Three Major Control Measures"

April 2, 2020

1 Mathematical model and configuration

The mathematical model used in this article is derived from the susceptible-exposed-infectious-removed (SEIR) [7, 6, 2, 8] augmented to incorporate human movement, and separate the infected compartment into reported cases and unreported cases. The model is illustrated in the flowchat below:

In the model, for each state, the population is divided into six compartments: S_i(susceptible), E_i(latent), I_i(reported infections), U_i(unreported infections) and R_i(removed: either recovered or deceased individuals). The subindex i stands for the index for the state. Among the six compartments, S, E and U are "free" people and can move from state to state, while I, and R are monitored and isolated.

The model is composed of 51 coupled ordinary differential equation (ODE) systems. For each

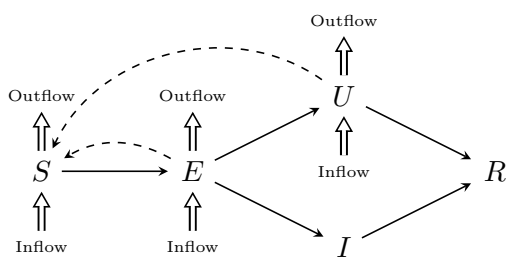


Figure S1: Illustration of the travel flow-network augmented susceptible-exposed-infectious-removed model.

state i, the model writes as:

$$\begin{cases} \dfrac{\mathrm{d}S_i}{\mathrm{d}t} = -\dfrac{b_i S_i(U_i+\gamma E_i)}{P_i} + \sum\limits_{j\neq i}\alpha_t n_{ij}\dfrac{S_j}{P_j} - \sum\limits_{j\neq i}\alpha_t n_{ji}\dfrac{S_i}{P_i} \\ \dfrac{\mathrm{d}E_i}{\mathrm{d}t} = \dfrac{b_i S_i(U_i+\gamma E_i)}{P_i} - \dfrac{E_i}{D_e} + \sum\limits_{j\neq i}\alpha_t n_{ij}\dfrac{E_j}{P_j} - \sum\limits_{j\neq i}\alpha_t n_{ji}\dfrac{E_i}{P_i} \\ \dfrac{\mathrm{d}I_i}{\mathrm{d}t} = r_i\dfrac{E_i}{D_e} - c_I\dfrac{I_i}{D_c} - (1-c_I)\dfrac{I_i}{D_l} \\ \dfrac{\mathrm{d}U_i}{\mathrm{d}t} = (1-r_i)\dfrac{E_i}{D_e} - c_U\dfrac{U_i}{D_c} - (1-c_U)\dfrac{U_i}{D_l} + \sum\limits_{j\neq i}\alpha_t n_{ij}\dfrac{U_j}{P_j} - \sum\limits_{j\neq i}\alpha_t n_{ji}\dfrac{U_i}{P_i} \\ \dfrac{\mathrm{d}R_i}{\mathrm{d}t} = c_I\dfrac{I_i}{D_c} + (1-c_I)\dfrac{I_i}{D_l} + c_U\dfrac{U_i}{D_c} + (1-c_U)\dfrac{U_i}{D_l} \end{cases} . \tag{1}$$

The ODE system is equipped with the following initial data ($t = 0$ standing for March 1, 2020):

$$S_i(0) = N_i - E_{i0} - U_{i0} - I_{i0}\,, \quad E_i(0) = E_{i0}\,, \quad I_i(0) = I_{i0}\,, \quad U_i(0) = U_{i0}\,, \quad R_i(0) = 0. \tag{2}$$

In the equation, the unit for t is one day. $N_i(t)$ is the total population of state i at time t, and $P_i = S_i + E_i + U_i$ is the free population. n_{ij} is the number of inflow from state j to state i. b_i and r_i are the transmission rate and reporting rate of state i. c_I (c_U, resp.) is the proportion of positive cases that show critical condition for I (unreported cases U, resp.). D_e is the latent period. D_c and D_l are the infectious periods of critical cases and mild cases. α_t is a parameter to tune the traffic flow.

We emphasize two main differences in modeling compared with literature. In [10], the authors study the inter-city traffic and its impact on the spreading of COVID-19 in China. The situation in China and that in the US are very different. In China, the epicenter is clear: the city of Wuhan, Hubei province, and the outbreak starts mid-January, 2020. The COVID-19 outbreak in the US, however, is multi-sourced. The consequence is that in the model in [10], the initial condition for cities excepts Wuhan is clear: the latent, the reported and the unreported cases are all zero. In this model, however, the initial conditions E_{i0} are unclear for all states; Another big difference is, according to clinical findings, the latent cases also have the potential of transmitting the virus, and thus we add the interaction of E_i with S_i into the increment of E_i [9, 10, 1].

The unknown parameters and state variables in the equation set are

* b_i: the transmission rate with non-informative prior range $[1, 1.5]$;
* r_i: the report rate with non-informative prior range $[0.1, 0.3]$;
* E_{i0}: the data for the latent population with non-informative prior range $[0, 500]$.
* U_{i0}: the initial data for the unreported population with non-informative prior range $[0, 200]$.
* S_{i0}: the initial data for the susceptible population defined by $N_i - E_{i0} - I_{i0} - U_{i0}$.

Other parameters are:

γ: the transmission ratio between unreported and latent. In the simulation we set it to be 0.5;

D_c: the average duration of infection for critical cases. We assume $D_c = 2.3$ days [5].

D_e: the average latent period. According to [13], $D_e = 5.2$ days.

D_l: the average duration of infection for mild cases. We assume $D_l = 6$ days.

α_t: the ratio of interstate travel volume compared to that of 2019 during the same period. The travel flow information n_{ij} was extracted from the SafeGraph mobility data, and we set $\alpha_t = 0.5$ to represent the travel reduction situation observed in the year of 2020.

c_I: proportion of critical cases among all reported cases. We assume $c_I = 0.1$.

c_U: proportion of critical cases among all unreported cases. We assume $c_U = 0.2$.

There is an essential assumption made in the model: the homogeneity in the population. It means that the traffic flow is a good representation of the total population without considering their demographic and socioeconomic characteristics. The susceptible, exposed, and unreported move in and out of states at the same rate. This explains the $\frac{S_i}{P_i}$, $\frac{E_i}{P_i}$ and $\frac{U_i}{P_i}$ terms in the $S_i/E_i/U_i$ equation.

The effective reproductive number R_e could be computed as

$$R_e = \frac{b}{E+U}\left[\gamma D_e E + \frac{D_c D_l U}{c_U D_l + (1-c_U) D_c}\right]. \tag{3}$$

R_e depends on time due to the time dependence of E and U.

2 Data acquisition

2.1 COVID-19 Observation Data

The only available data in the $(S-E-I-U-R)$ system, for every state, is I, the reported confirmed cases. We fetch the data from a community-developed, open source project: Novel Coronavirus (COVID-19) Cases, developed by the Center For Systems Science and Engineering at the Johns Hopkins University [3] [1].

2.2 Population and Human Mobility Data

We downloaded the total population data by state in 2019 from the US Census Bureau. In addition, we collected over 3.6 million points of interest (POIs) with travel patterns in the United States from the SafeGraph business venue database[2]. The SafeGraph's data sampling correlated highly with the United States Census populations[3]. These mobile location data consist of "pings" identifying the coordinates of a smartphone at a moment in time. To enhance privacy, SafeGraph excludes census block group (CBG) information if fewer than five devices visited a place in a month from a given CBG. For each POI, the records of aggregated visitor patterns illustrate the number of unique visitors and the number of total visits to each venue during the specified time window (i.e., March 1st to March 31st 2019 in our dataset), which could reflect the attractiveness of each venue and the national spatial interaction patterns during the last March travel. According to the ODE modeling needs, we further aggregated the travel patterns to the state-to-state spatial scale as shown in Figure S2. In the model, we set the parameter $\alpha_t = 0.5$ to represent the travel reduction situation observed in the year of 2020 [14].

[1] `https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv`

[2] `www.safegraph.com`

[3] `www.safegraph.com/blog/what-about-bias-in-the-safegraph-dataset`

Figure S2: The map of US state-to-state travel patterns aggregated in March 2019 (Data source: SafeGraph; Visualization: kepler.gl)

.

3 Parameter fitting methodology

Each state has its own $S-E-I-U-R$ data. We assume all the states have their own transmission rate b, and the reporting rate r. In this section we discuss the method we apply to recover these parameters.

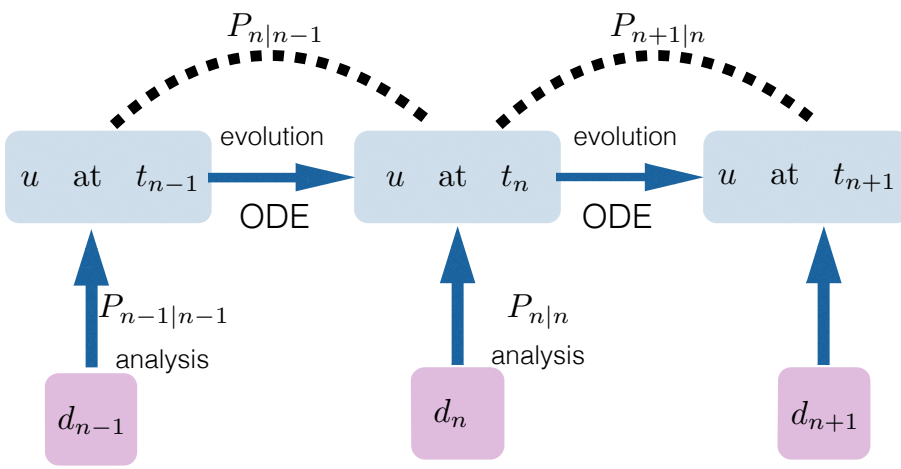
To identify the parameters is a typical data assimilation problem: one has the knowledge from an underlying ODE model, and the access to evolution data. The goal is to build a probability density function that reveals the possible value and the probability of the state variable. Two main ingredients in DA are ODE simulation, and the Bayesian analysis. The ODE system serves as the prior information, and the Bayesian formula blends such dynamical system with the newly fetched data to generate a posterior distribution of the state variables and the parameters. The general flow chart of data assimilation is found in Figure S3.

In the flowchart, $P_{n|n-1}$ is the probability density of u upon the evolution step, and $P_{n|n}$ is the probability density obtained through the analysis step.

In our case, u is the augmented state variable that includes both the unknown parameters (b, r) and the state variable (I, E, S, U, R) for every state:

$$u = (b_i, r_i, I_i, E_i, S_i, U_i, R_i)^\top \in \mathbb{R}^{7k}\,,$$

where k is the number of states/regions used in the fitting. We transpose it to make it into a column vector. We are aiming at building a distribution density function of u over the $\mathbb{R}^{7k}$ space. $\mathcal{G}_{m,n}(u)$ is the solution to the ODE (1) at time t_n with u serving as the parameter in the equation



evolution: $P_{n|n-1}(u) = \delta(u - \mathcal{G}_{n-1,n}(u'))P_{n-1|n-1}(u')$

analysis: $P_{n|n}(u) \propto P_{n|n-1}(u)q(d|u)$

Figure S3: Flowchart of data assimilation: u is the state variable, $\mathcal{G}_{n-1,n}(u)$ is the forward map by running the ODE from time step t_{n-1} to t_n. q is the likelihood function that measures the probability of the error term $d - \mathcal{M}u$ where $\mathcal{M}$ is the measuring operator.

at time t_m. The measuring operator is $\mathcal{M}(u) = [I_i]_{i=1}^k$, and thus we denote

$$\mathcal{M}(u) = M \cdot u\,, \quad \text{with} \quad M = \otimes^k[0,0,1,0,0,0]]\,,$$

meaning M is a $7k$ length vector that has k nontrivial entries that pick up all I_i information.

We further assume the collected data has a Gaussian perturbation from the true measuring operator:

$$d_n = \mathcal{M}(u_n) + \eta\,, \quad \text{with} \quad \eta \sim \mathcal{N}(0, \sigma^2)\,,$$

and thus naturally the likelihood function is:

$$q(d|u) = \frac{1}{2\pi\sigma} e^{-\frac{|d-\mathcal{M}(u)|^2}{2\sigma^2}}\,.$$

Then together with the two formula for the evolution step and the analysis step:

- Evolution: $P_{n|n-1}(u) = \delta(u - \mathcal{G}_{n-1,n}(u'))P_{n-1|n-1}(u')$
- Analysis: $P_{n|n}(u) \propto P_{n|n-1}(u)q(d|u)$

one can iteratively update $P_{n|n}(u)$, giving $P_{n-1|n-1}$, the probability density of u at time t_{n-1}. In the equation $\propto$ is the proportional sign: one needs to normalize $P_{n|n}$ to make it a probability density function so that $\int P_{n|n}(u)\mathrm{d}u = 1$.

There are many choices of data assimilation methods. We choose to utilize Ensemble Kalman Filter that is steered towards analyzing systems having high dimensional state variables. It is a technique, derived from the classical Kalman Filter for the application in atmospheric science, with the analytical covariance matrix replaced by the ensemble version, eliminating the computation of the the Kalman gain matrix and the Riccati Equation that are typically expensive in high dimensional space. It is proved to the effective in the Gaussian case for linear forward model

and the measuring operator [4, 11, 12]. The idea is to sample a fixed number of particles on the state variable space according to the initial distribution, and move these particles around at every discrete time step with certain dynamics, to represent the newly adjusted distribution. Denote the number of particles by N, and the j-th particle after the evolution step at time t_n by $u_n^{j,\mathrm{E}}$, and the j-th particle after the analysis step at t_n by $u_n^{j,\mathrm{A}}$, we now summarize the algorithm:

- Evolution:
$$u_n^{j,\mathrm{E}} = FE(\mathcal{G}_{n-1,n}(u_{n-1}^{j,\mathrm{A}}))\,.$$
where FE is the Forward-Euler discretization applied on ODE (1).

- Analysis:
$$u_n^{j,\mathrm{A}} = u_n^{j,\mathrm{E}} + \mathrm{Cov}_n^{\mathrm{up}}\left(\mathrm{Cov}_n^{\mathrm{pp}} + \sigma^2\right)^{-1}\left(d_n + \xi_n^j - \mathcal{M}u^{j,\mathrm{E}}\right),$$
where ξ_n^j is a k-length vector with each entry being random variable i.i.d. drawn from $\mathcal{N}(0,\sigma^2)$, d_n is the k-length vector collecting the reported infected data on k counties. To compute the covariance matrices, we set:
$$\bar{u}_n^{\mathrm{E}} = \frac{1}{N}\sum_{j=1}^{N} u_n^{j,\mathrm{E}}\,, \quad \mathrm{Cov}_n^{\mathrm{uu}} = \frac{1}{N}\sum_{j=1}^{N}(u_n^{j,\mathrm{E}} - \bar{u}_n^{\mathrm{E}})(u_n^{j,\mathrm{E}} - \bar{u}_n^{\mathrm{E}})^\top\,.$$
Then naturally
$$\mathrm{Cov}_n^{\mathrm{up}} = \mathrm{Cov}_n^{\mathrm{uu}} \cdot M^\top\,, \quad \mathrm{Cov}_n^{\mathrm{up}} = M \cdot \mathrm{Cov}_n^{\mathrm{uu}} \cdot M^\top\,.$$
are matrices of size $7k \times k$ and $k \times k$ respectively.

4 Results

The results are divided into two categories: 1. parameter fitting; 2. COVID-19 infection prediction. The computation is done on the state level. Data from 50 states and D.C. (thus $k = 51$) in the United States are used. The model and data assimilation analyses were ran from March 1 to March 20, 2020, and we predict the future infectious cases in different states from March 21 to April 29, 2020 with different parameter setting scenarios.

4.1 State results

For parameter fitting, we utilize the method discussed in Section 3. Total 2000 samples with non-informative prior are adopted to determine $7 \times 51 = 357$ state variables. The standard deviation of noise is set to be $\sigma = 10$. In Figure S4-S5, we plot the susceptible, exposed, unreported infections and removed, in time, for the top 10 states with most total confirmed cases as of March 20, respectively. In Figure S6-S7, we plot the reported infections in time, for these 10 states. For most states, the number of reported case grows essentially exponentially fast. Figure S8-S11 show the inferred transmission rate b and reported rate r. Figure S12-S13 show the time series of effective reproductive number R_e for different states. The signal in the data is rather weak, and for some states, the number of E and U cannot be inferred at the early stage of the breakout, leading to $R_e = 0$ for a small period of time for some states.

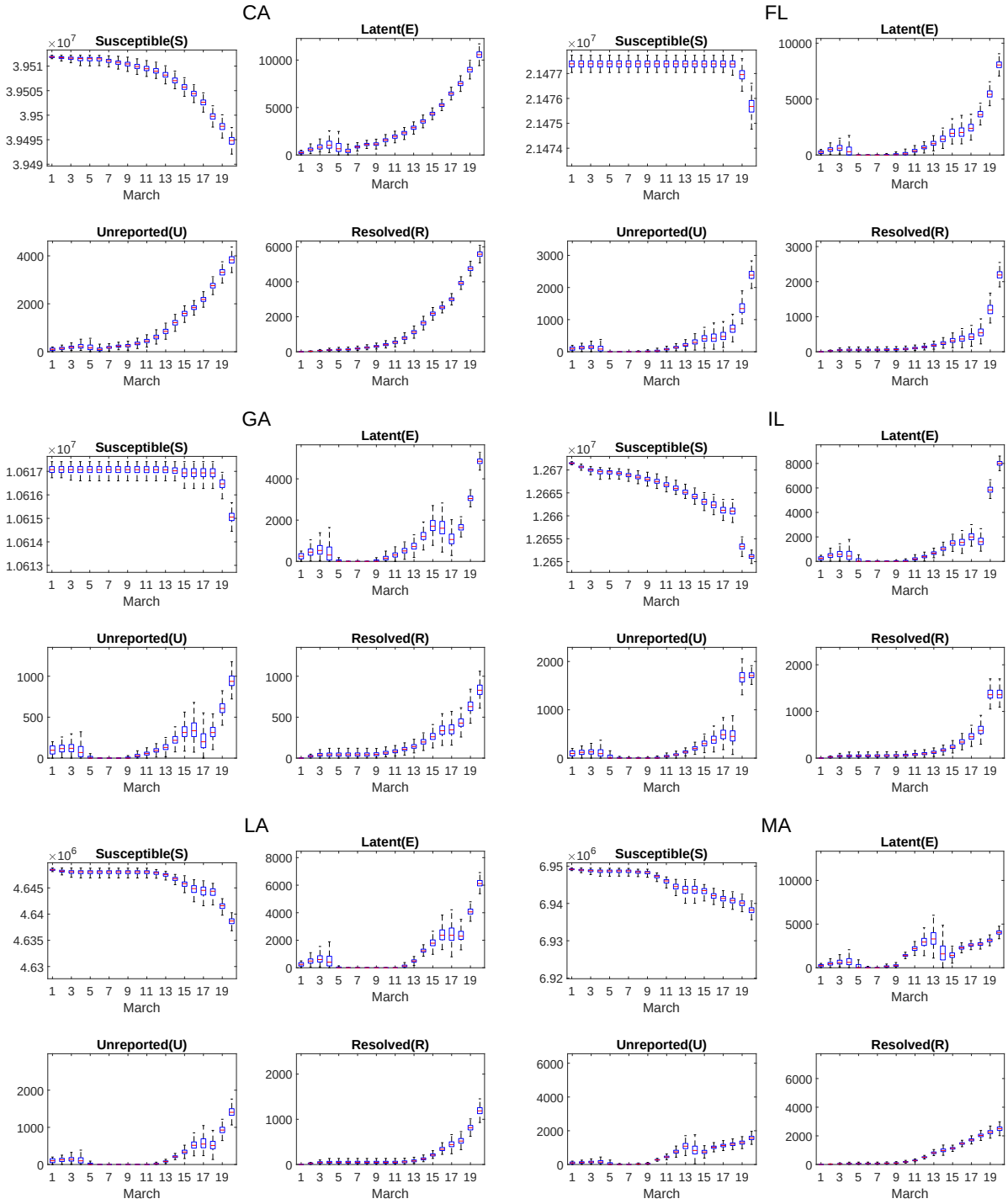


Figure S4: Simulation of $S-E-U-R$ for different states. The box and whiskers show the median, interquartile range, and 95% credible intervals.

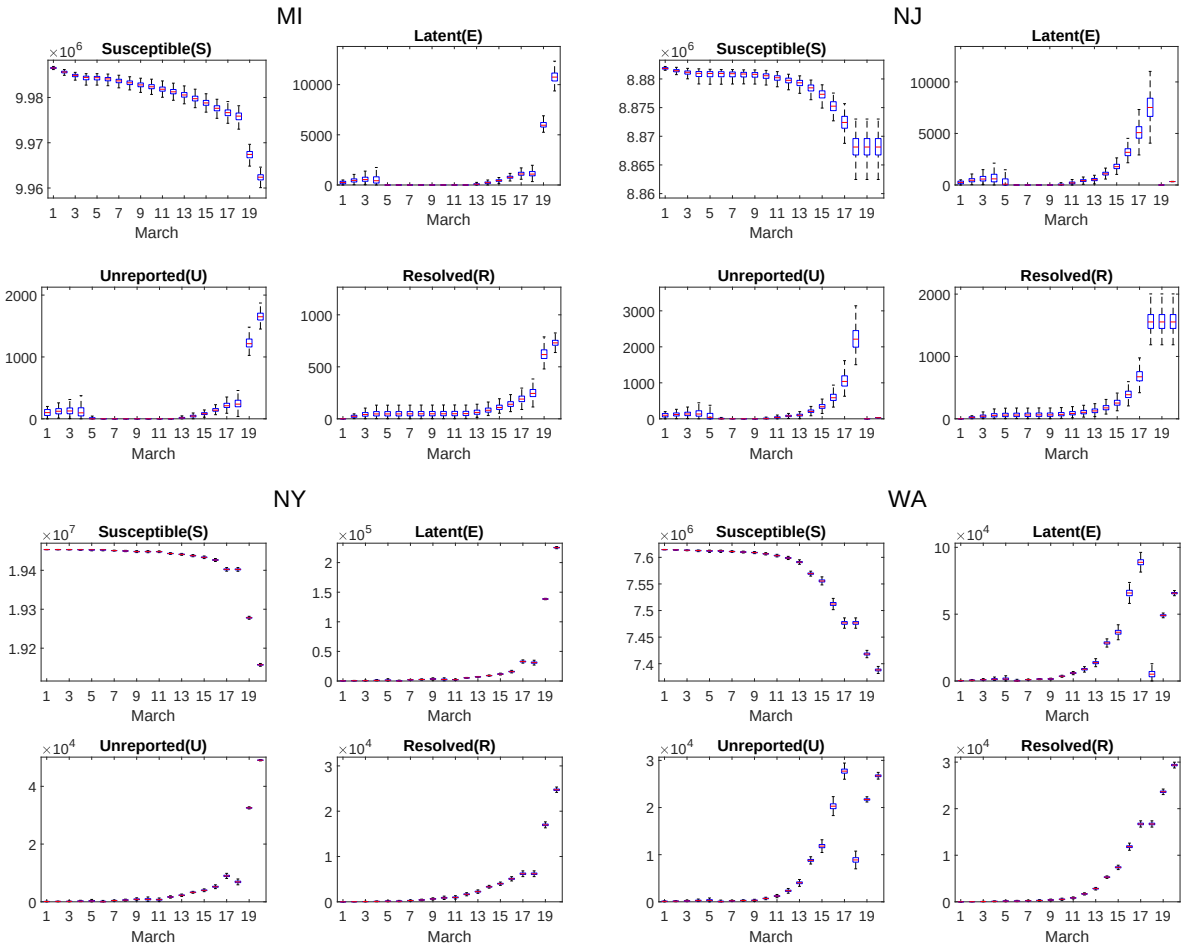


Figure S5: Simulation of $S - E - U - R$ for different states. (continued)

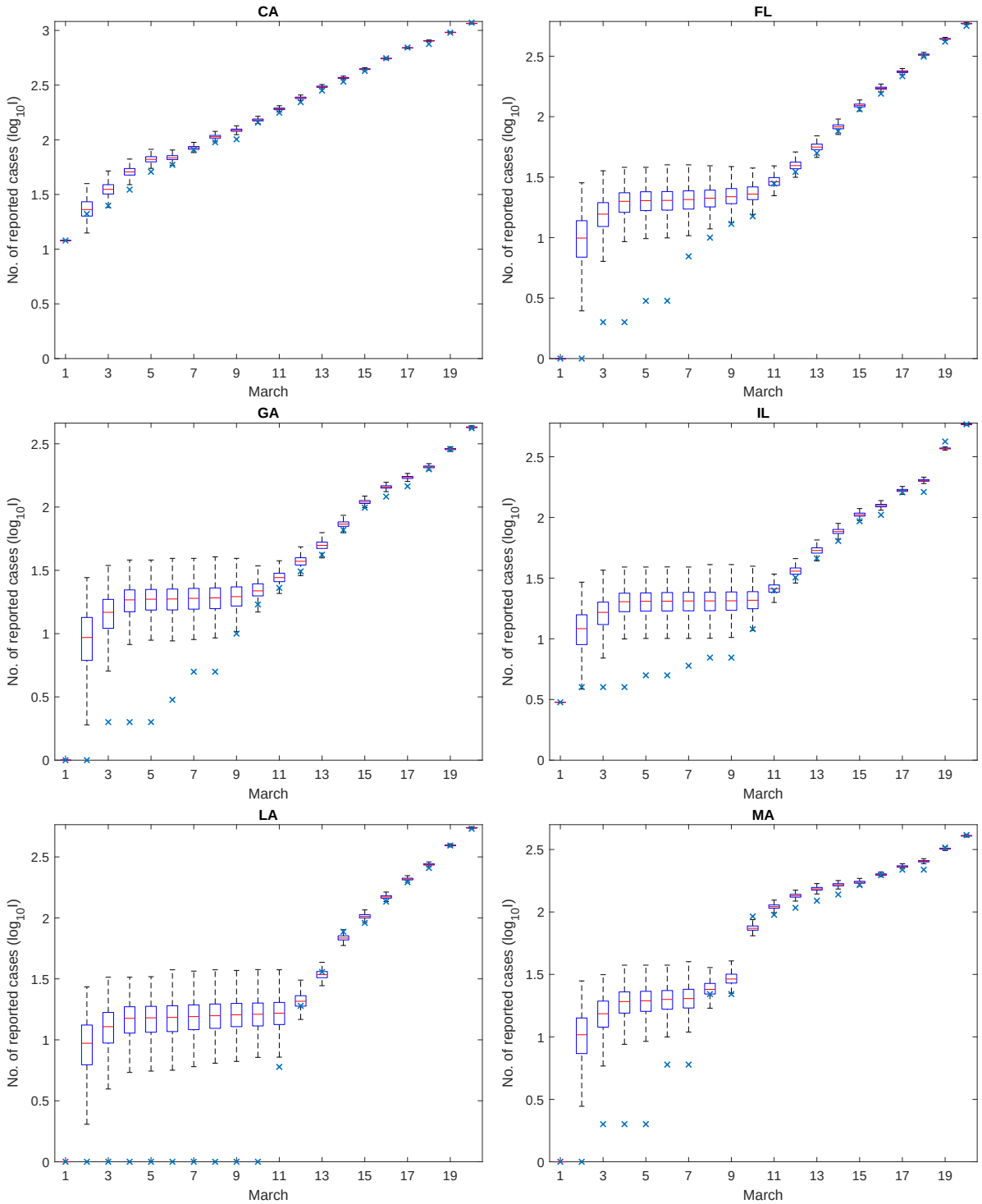


Figure S6: Simulation of confirmed cases I (boxes) and the true confirmed cases I_{true} (blue x's) for different states. The box and whiskers show the median, interquartile range, and 95% credible intervals.

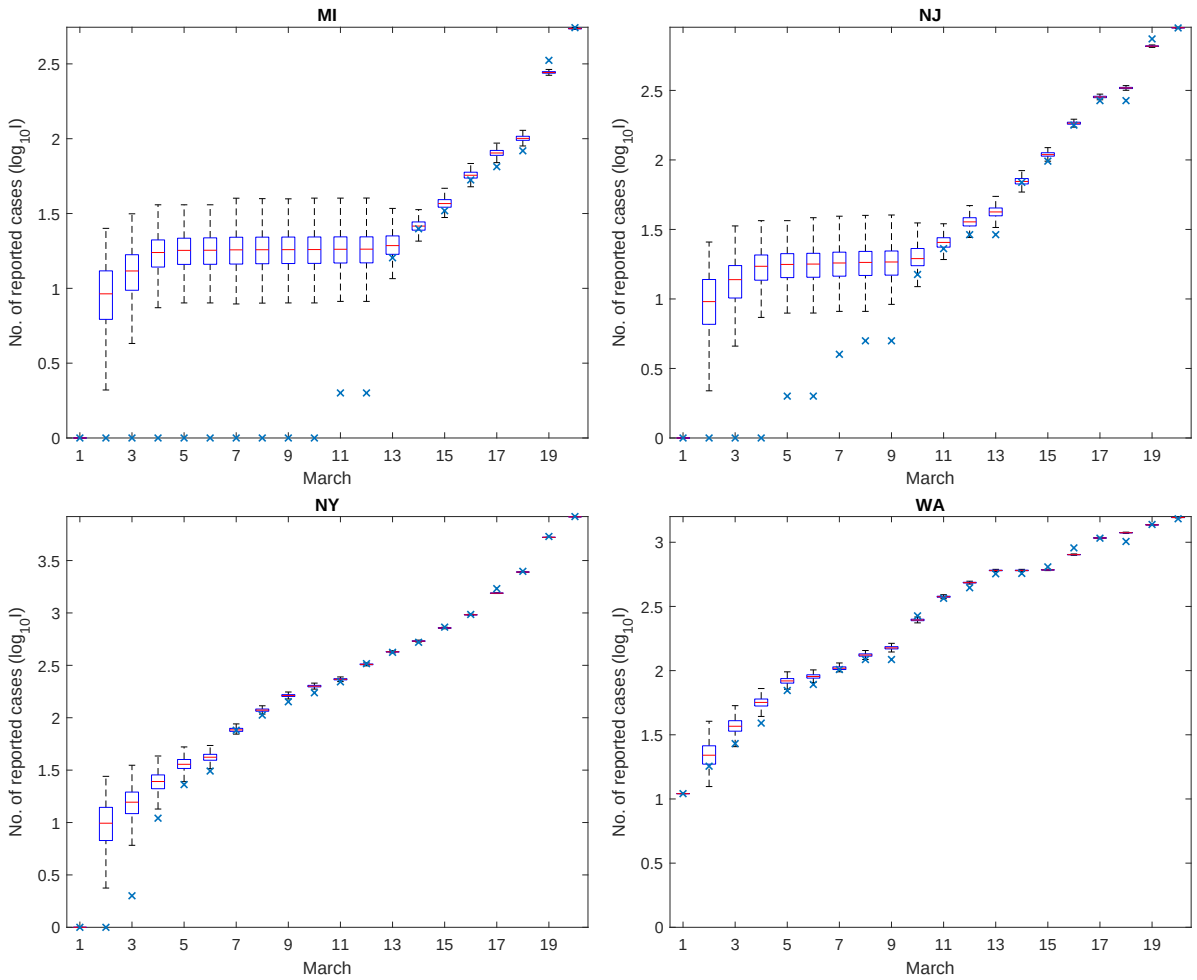


Figure S7: Simulation of confirmed cases I (boxes) and the true confirmed cases I_{true} (blue x's) for different states. (continued)

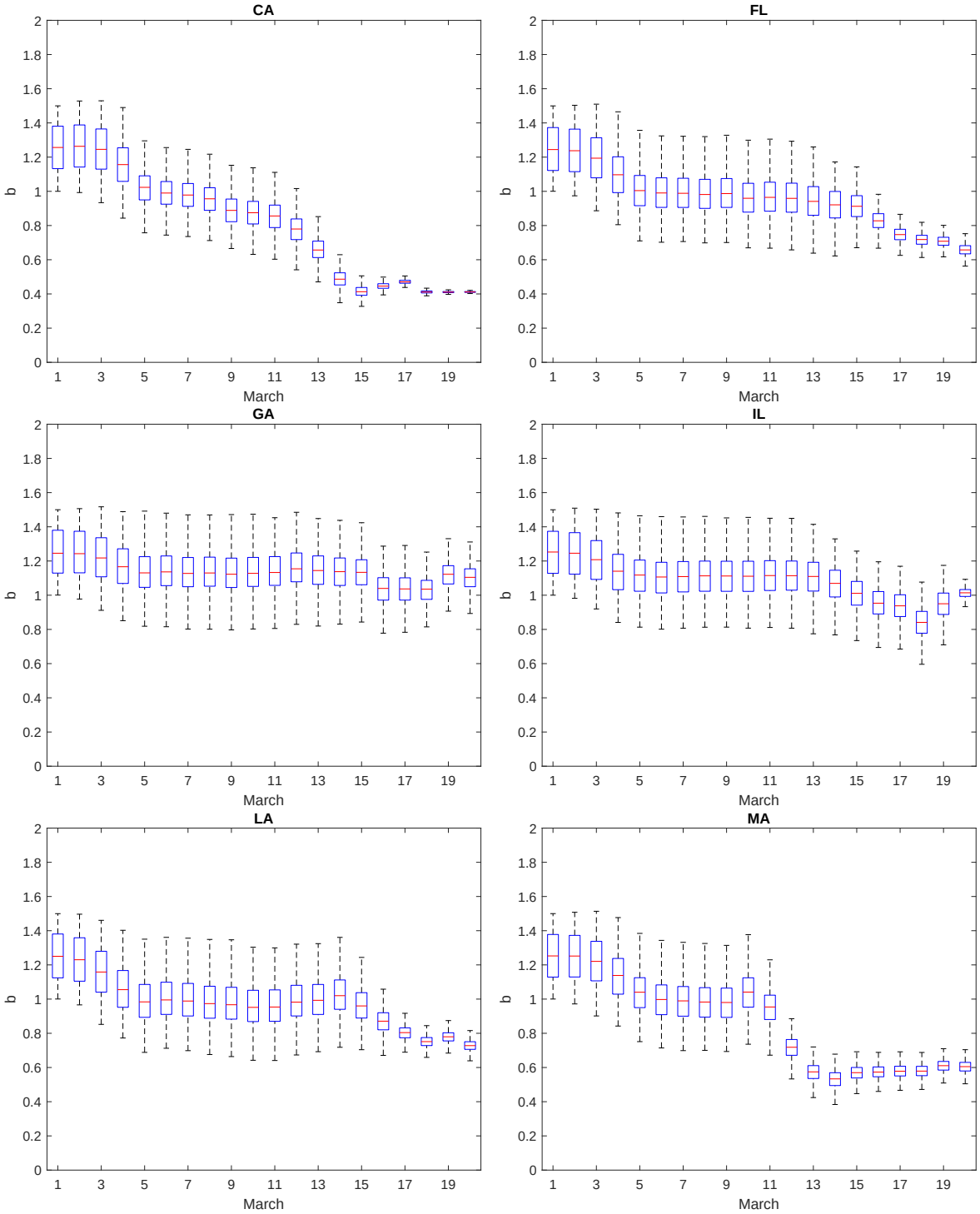


Figure S8: Inferred transmission rate b for different states. The box and whiskers show the median, interquartile range, and 95% credible intervals.

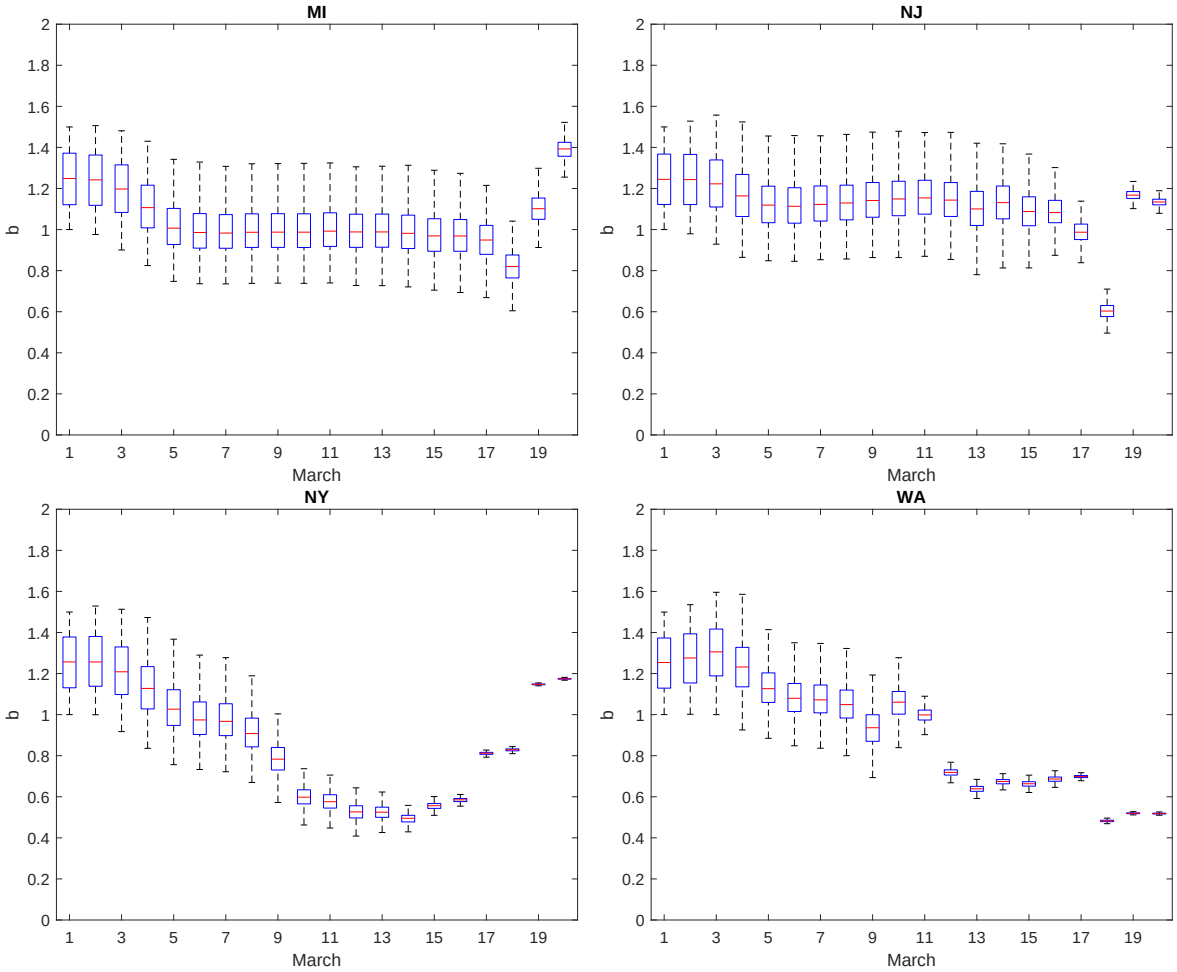


Figure S9: Inferred transmission rate b for different states. (continued)

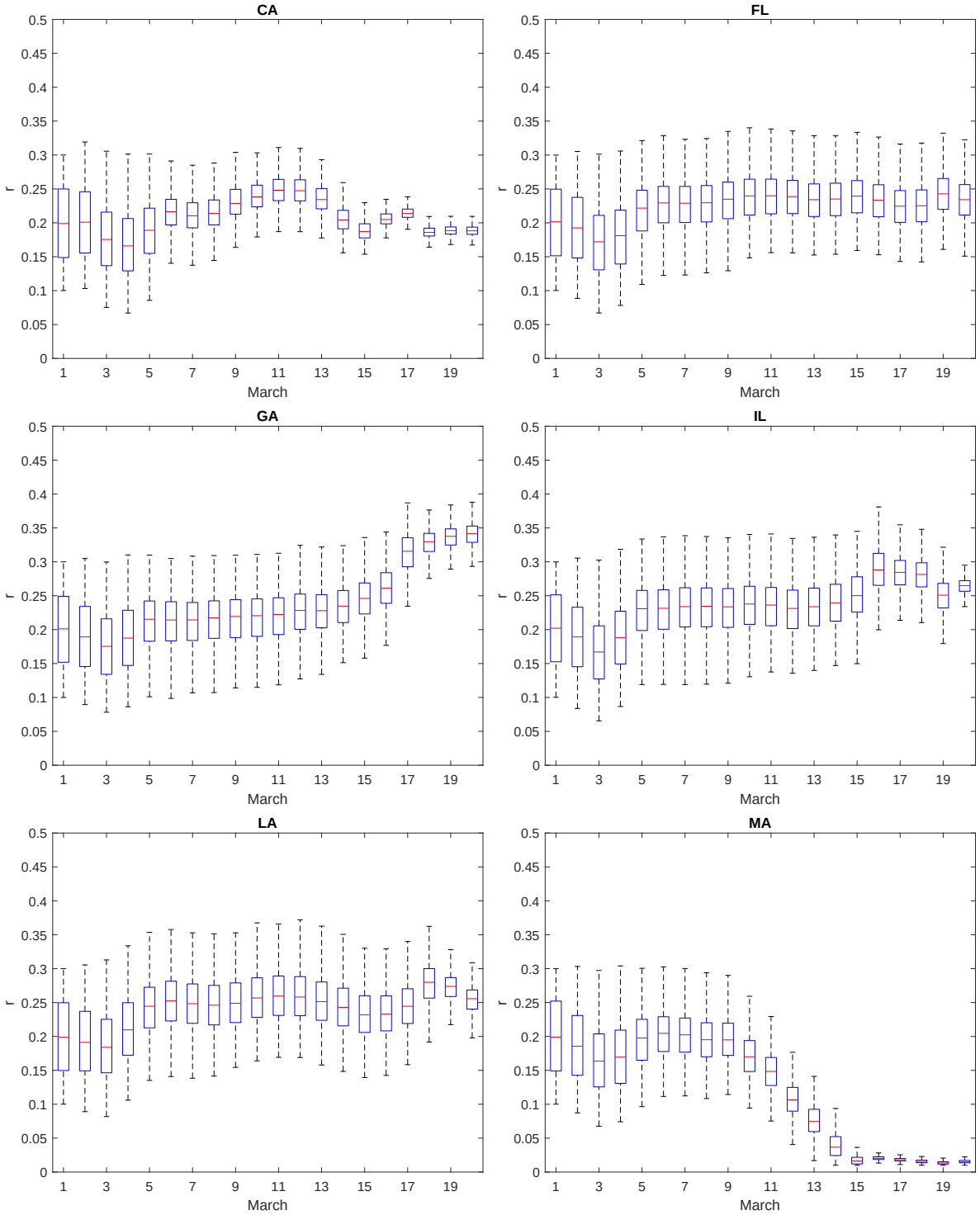


Figure S10: Inferred reporting rate r for different states. The box and whiskers show the median, interquartile range, and 95% credible intervals.

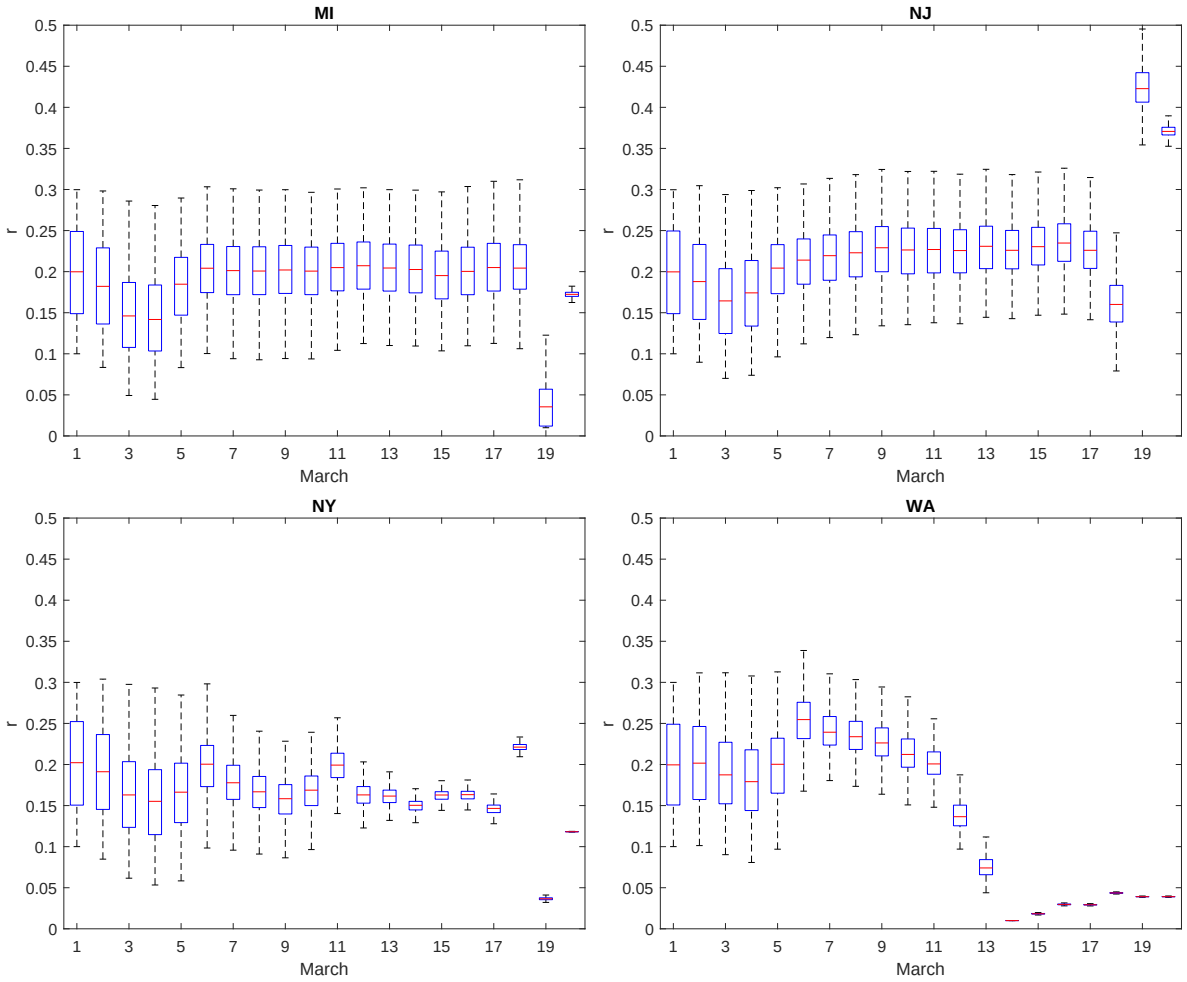


Figure S11: Inferred reporting rate r for different states. (continued)

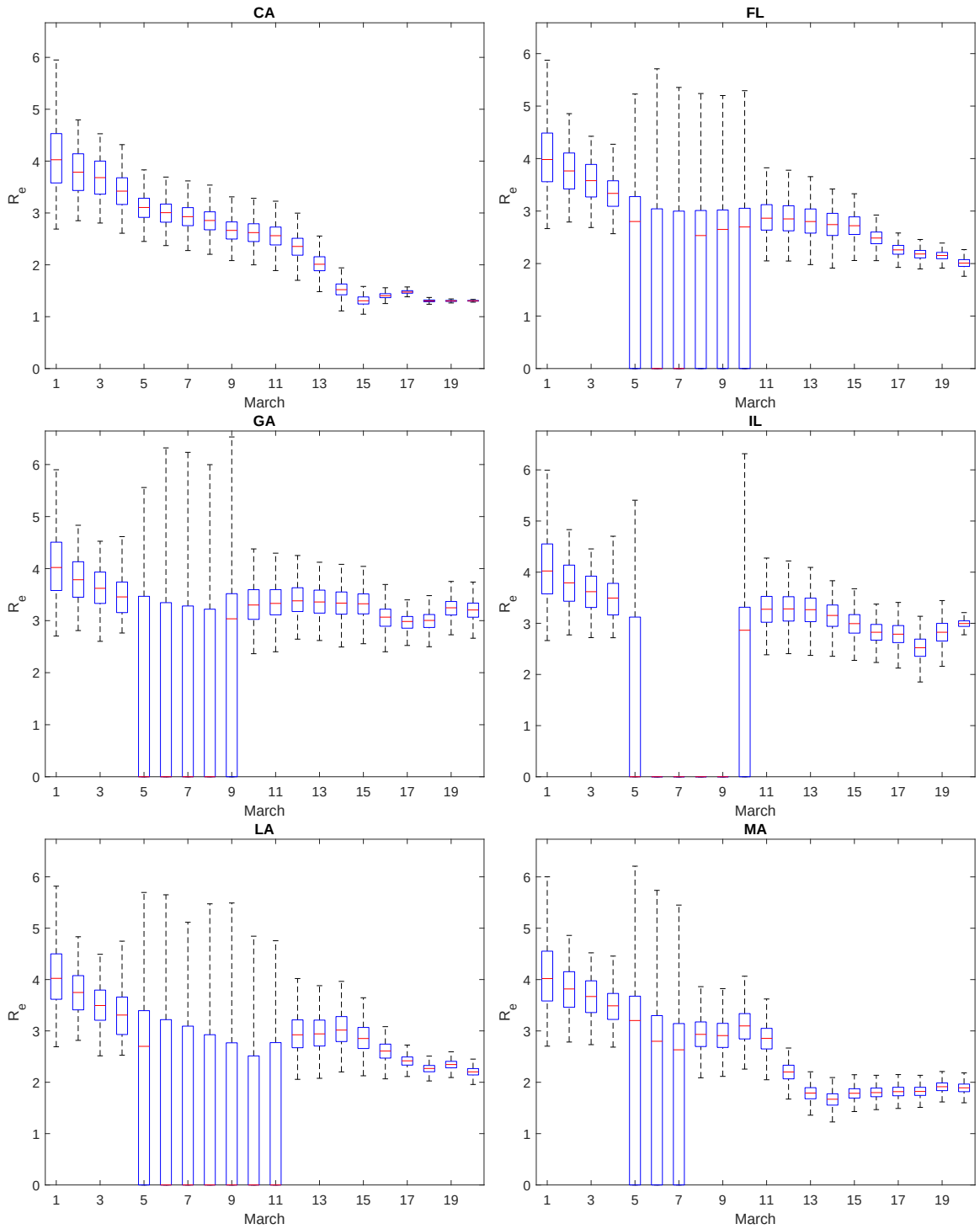


Figure S12: Inferred effective reproductive number R_e for different states. The box and whiskers show the median, interquartile range, and 95% credible intervals.

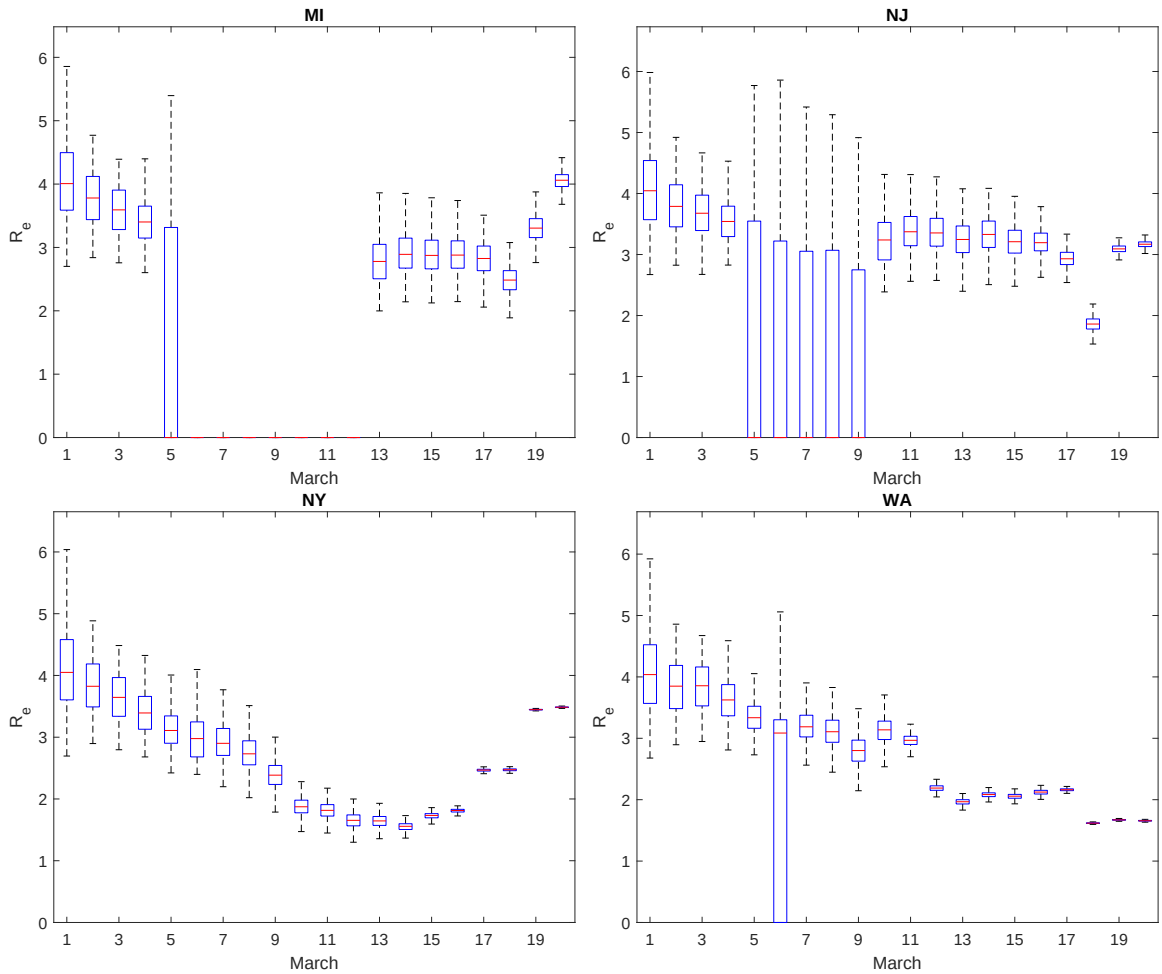


Figure S13: Inferred effective reproductive number R_e for different states. (continued)

4.2 Prediction

The first part of our prediction includes the case study for different transmission rate b and reporting rate r. In the following, we define the ratio between transmission rate in prediction step and data assimilation step (March 20) by α_b, and define the ratio between unreporting rate $1-r$ in prediction step and data assimilation step (March 20) by α_r, namely

$$b_i = \alpha_b b_{i,0}\,, \quad r_i = 1 - \alpha_r(1 - r_{i,0})$$

Fig S15 shows the effective reproductive number R_e on April 29 as a function of α_r and α_b for five states.

We further take a proactive approach by directly identify and quarantine the exposed population. The time in days used to discover and isolate them is, on average D_q. The model is changed to:

$$\begin{cases} \dfrac{\mathrm{d}S_i}{\mathrm{d}t} = -\dfrac{b_i S_i(U_i + \gamma E_i)}{P_i} + \sum\limits_{j\neq i} \alpha_t n_{ij} \dfrac{S_j}{P_j} - \sum\limits_{j\neq i} \alpha_t n_{ji} \dfrac{S_i}{P_i} \\ \dfrac{\mathrm{d}E_i}{\mathrm{d}t} = \dfrac{b_i S_i(U_i + \gamma E_i)}{P_i} - \dfrac{E_i}{D_q} + \sum\limits_{j\neq i} \alpha_t n_{ij} \dfrac{E_j}{P_j} - \sum\limits_{j\neq i} \alpha_t n_{ji} \dfrac{E_i}{P_i} \\ \dfrac{\mathrm{d}I_i}{\mathrm{d}t} = -c_I \dfrac{I_i}{D_c} - (1 - c_I)\dfrac{I_i}{D_l} \\ \dfrac{\mathrm{d}U_i}{\mathrm{d}t} = -c_U \dfrac{U_i}{D_c} - (1 - c_U)\dfrac{U_i}{D_l} + \sum\limits_{j\neq i} \alpha_t n_{ij} \dfrac{U_j}{P_j} - \sum\limits_{j\neq i} \alpha_t n_{ji} \dfrac{U_i}{P_i} \\ \dfrac{\mathrm{d}R_i}{\mathrm{d}t} = c_I \dfrac{I_i}{D_c} + (1 - c_I)\dfrac{I_i}{D_l} + c_U \dfrac{U_i}{D_c} + (1 - c_U)\dfrac{U_i}{D_l} \\ \dfrac{\mathrm{d}Q_i}{\mathrm{d}t} = \dfrac{E_i}{D_q} \end{cases} . \tag{4}$$

As shown in Fig. 3c, the timely self-quarantine and strict isolation right away or starting at most about 3.6 days (the median value) after exposed to the SARS-CoV-2 virus is found most effective in containing the COVID-19 outbreak for most contagious states in the US.

4.3 Limitation

For the scope of this paper, we do not consider the hospital capacity, and thus we assume the current or timely built facility can accommodate all reported infected cases. However, if one can have access to the hospital capacity data in each state, the capacity information can be fed to the model as well. Furthermore, we do not distinguish mild and severe symptoms in the model.

References

[1] CNN Report, Infected people without symptoms might be driving the spread of coronavirus more than we realized, available at https://www.cnn.com/2020/03/14/health/coronavirus-asymptomatic-spread/index.html.

[2] F. Brauer, *Compartmental models in epidemiology*, in Mathematical epidemiology, Springer, 2008, pp. 19–79.

[3] E. Dong, H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time*, The Lancet Infectious Diseases.

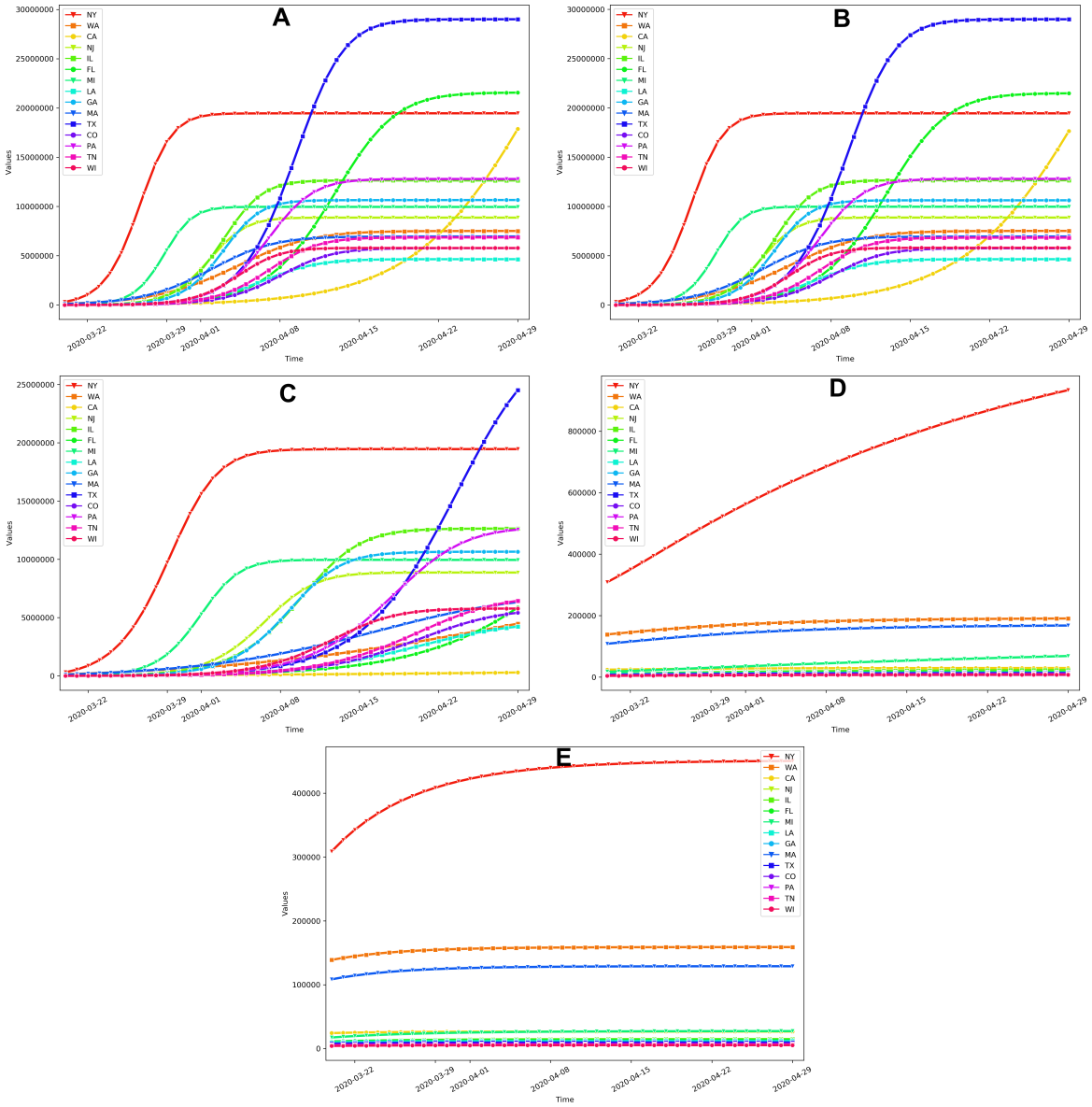


Figure S14: The prediction time series of the total infected population in the 15 most affected states under five scenarios: (A) $\alpha_r = 1$ and $\alpha_b = 1$, i.e., all parameters took the values of the initial configuration, obtained through data assimilation method using the numbers of confirmed cases during March 1 – March 20, 2020; (B) the travel flow was reduced to $\alpha_t = 0.05$, while other parameters values remained unchanged; (C) $\alpha_r = 0.1$ and $\alpha_b = 1$; (D) $\alpha_r = 1$ and $\alpha_b = 0.1$; (E) $\alpha_r = 0.1$, $\alpha_b = 0.1$. In the simulations, the transmission rate was set to be $b = \alpha_b b_0$ and the reporting rate was set to be $r = 1 - \alpha_r(1 - r_0)$. Where r_0 and b_0 were the reporting rate and the transmission rate on March 20, 2020, which are inferred from the data assimilation step.

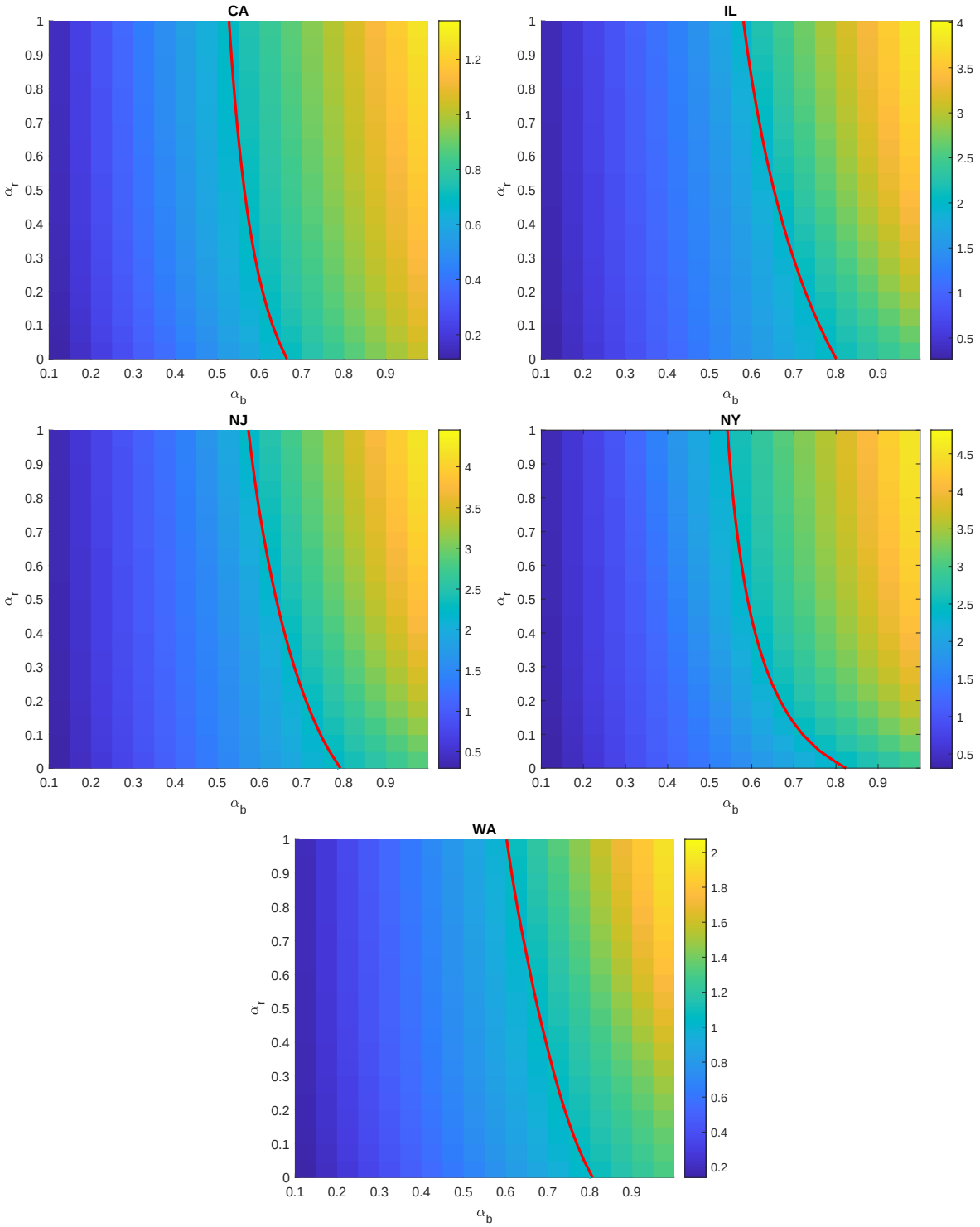


Figure S15: R_e on April 29 for different α_b and α_r. The red line is the level set $R_e = 1$. It can be seen that increasing the reported rate helps to diminish the reproductive number.

[4] G. Evensen, *The ensemble kalman filter for combined state and parameter estimation*, IEEE Control Systems Magazine, 29 (2009), pp. 83–104.

[5] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, et al., *Clinical characteristics of coronavirus disease 2019 in China*, New England Journal of Medicine, (2020).

[6] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM review, 42 (2000), pp. 599–653.

[7] W. O. Kermack and A. 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 (1927), pp. 700–721.

[8] S. Lai, N. W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J. R. Floyd, A. Wesolowski, C. Zhang, X. Du, H. Yu, et al., *Effect of non-pharmaceutical interventions for containing the COVID-19 outbreak: an observational and modelling study*, medRxiv, (2020).

[9] N. H. Leung, D. K. Chu, E. Y. Shiu, K.-H. Chan, J. J. McDevitt, B. J. Hau, H.-L. Yen, Y. Li, D. KM, J. Ip, et al., *Respiratory virus shedding in exhaled breath and efficacy of face masks*, Nature Medicine, (2020).

[10] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, *Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)*, Science, (2020).

[11] S. Reich and C. Cotter, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press, 2015.

[12] A. Stuart and K. Zygalakis, *Data assimilation: A mathematical introduction*, tech. rep., Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2015.

[13] C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, A. Pan, et al., *Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in Wuhan, China*, medRxiv, (2020).

[14] M. S. Warren and S. W. Skillman, *Mobility changes in response to COVID-19*, Descartes Labs, (2020).