Implications of delayed reopening in controlling the COVID-19 surge in Southern and West-Central USA

Raj Dandekar[1, 6, *], Emma Wang[2], George Barbastathis[3, 4], and Chris Rackauckas[5]

[1]Department of Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Electrical Engineering and Computer Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Singapore-MIT Alliance for Research and Technology (SMART) Centre, Singapore 138602
[5]Department of Applied Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Lead Contact
[*]Correspondence: rajd@mit.edu

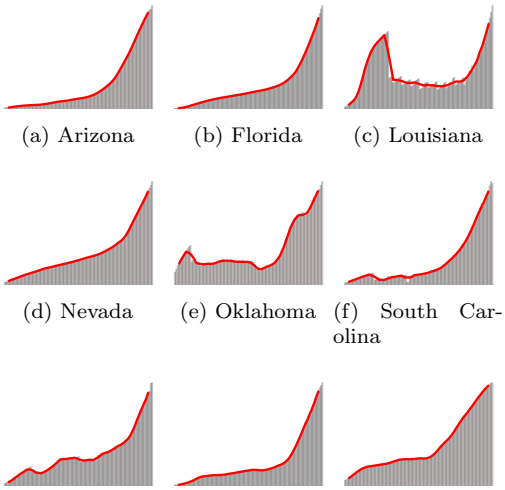
November 16, 2020

1 SUMMARY

In the wake of the rapid surge in the Covid-19 infected cases seen in Southern and West-Central USA in the period of June-July 2020, there is an urgent need to develop robust, data-driven models to quantify the effect which early reopening had on the infected case count increase. In particular, it is imperative to address the question: How many infected cases could have been prevented, had the worst affected states not reopened early? To address this question, we have developed a novel Covid-19 model by augmenting the classical SIR epidemiological model with a neural network module. The model decomposes the contribution of quarantine strength to the infection timeseries, allowing us to quantify the role of quarantine control and the associated reopening policies in the US states which showed a major surge in infections. We show that the upsurge in the infected cases seen in these states is strongly co-related with a drop in the quarantine/lockdown strength diagnosed by our model. Further, our results demonstrate that in the event of a stricter lockdown without early reopening, the number of active infected cases recorded on 14 July could have been reduced by more than 40% in all states considered, with the actual number of infections reduced being more than $100,000$ for the states of Florida and Texas. As we continue our fight against Covid-19, our proposed model can be used as a valuable asset to simulate the effect of several reopening strategies on the infected count evolution; for any region under consideration.

2 INTRODUCTION

The Coronavirus respiratory disease 2019 originating from the virus "SARS-CoV-2"[1,2] has led to a global pandemic, leading to more than 50 million confirmed global cases in more than 200 countries as of November 13, 2020.[3] In the United States, the first infections were detected in Washington State as early as January 20, 2020,[4] and now it is being reported that the virus had been circulating undetected in New York City as early as mid-February.[5] As of September 21,



(a) Arizona (b) Florida (c) Louisiana

(d) Nevada (e) Oklahoma (f) South Carolina

(g) Tennessee (h) Texas (i) Utah

Figure 1: Active infected cases over time as of July 14, 2020, shown with a 7-day moving average, for the Southern and West-Central states considered in the present study.

2020, the United States has ≈ 6.9 million infected cases since the virus began to spread.

Since the second week of June, a second surge of Covid-19 was seen in the United States,[6] with rapidly increasing daily infected cases, hospitalization rates and death rates.[7,8] Initially driven by disastrous situations in the states of Arizona, South Carolina, Texas, Florida and Georgia,[6] the surge in cases was also later seen in several other Southern and West-Central states.[9] This surge can be seen in figure 1 which shows the active infected cases over time as of July 14, 2020 with a 7-day moving average for 9 states. States which reopened early show a generally strong co-relation with the rise in the infected cases over the 3-month period from late April to mid July 2020.[9] For example, states which opened before May 15 showed daily infected case increments of: Florida (1393 %), Arizona (858 %), South Carolina (999 %), Alabama (547 %), Oklahoma (477 %), Tennessee (279 %), Georgia (245 %), Mississippi (215 %), Nevada (697 %), Texas (680 %) and Utah (287 %); while states which reopened after May 29 showed values of: Michigan (16 %), Pennsylvania (−26 %), New York (−52 %), New Jersey (−32 %) and Illinois (−54 %). Thus, although early reopening seems to be co-related to the second surge of cases seen in the USA, there is a need for robust, data-driven quantification of the effect of early reopening on the growth of infected count data. More importantly, it is of utmost importance to answer the question: How many infected cases could have been prevented, had the worst affected states not reopened early?

In an effort to address this question, we have developed a machine learning-aided epidemiological model. The novelty of our model arises from the fact that it allows us to decompose the contribution of quarantine/lockdown strength evolution to the infected data timeseries for the region under consideration. This enables us to simulate the effect of varying quarantine strength evolutions and hence varying reopening strategies on the infected count data. We define reopening as beginning when a state allows its stay-at-home order to expire, or, in the case of states that never issued a stay-at-home order, when a state first starts allowing non-essential businesses, such as dine-in restaurants and hair salons, to reopen.[10,11] Considering nine US states which showed a significant surge in cases since the last month, we demonstrate that our model shows a drop in the quarantine strength evolution when these states were reopened. Furthermore, we show that maintaining a strict lockdown without early reopening would have led to about $500,000$ fewer infected cases in all these states combined.

3 RESULTS

The first stage of our analysis is using our model,[12,13] called the QSIR model to diagnose the underlying quarantine strength evolution $Q(t)$ in the regions under consideration. By applying the QSIR model to more than 70 countries globally, we have established the validity of $Q(t)$ in accurately diagnosing the on-the-ground quarantine situation in majorly affected European, South American and Asian countries.[12] A slow growth of $Q(t)$ without a significant increase indicates relaxed quarantine policies, a sharp transition point in $Q(t)$ is indicative of a sudden ramp-up of quarantine measures, and an inflection point corresponds to the time when the quarantine response was the most rapid in the region under consideration. The results of our model applied globally to all continents are hosted publicly at covid19ml.org.

In this study, to perform the quarantine diagnosis to analyze the implications of delayed reopening, we applied the QSIR model to 9 US states which showed a significant surge in the infected case count in the last month: Arizona, Florida, Louisiana, Nevada, Oklahoma, South Carolina, Tennessee, Texas and Utah. Figure 2 shows representative results for Arizona, Nevada, South Carolina and Tennessee. The plots for the remaining states are provided in the Supplementary Information. Figures 2 a, d, g, j show the comparison of the infected and recovered count estimated by our model with the actual data. A reasonable agreement is seen for all states, with the model being able to capture the rise in infections seen in the tail end of the timeseries. The QSIR model details are provided in the Methods section; Mean Absolute Percentage Error (MAPE) values for the model along with the epochs required for convergence for each state are provided in Supplementary Information.

Figures 2 b, e, h, k show the quarantine strength evolution $Q(t)$ as learnt by the neural network module, which shows a decline whose starting point corresponds well to the time when these states began reopening[10,11] (shown by the green dotted line in the figures). In some states, the decline in $Q(t)$ starts later than the reopening date; possibly because of the time delay for population level changes to be seen in the infected count evolution, after reopening. $Q(t)$ trained by our model shows a significant drop after early reopening in all Southern and West-Central states that showed a surge in cases last month; whereas the North-Eastern states of New York, New Jersey and Illinois, which reopened late and showed no surge in infections, did not show a drop in $Q(t)$ (Table 1 and figures in Supplementary Information). Thus, the upsurge in the infected cases seen in these states is strongly co-related with a drop in the quarantine/lockdown strength $Q(t)$ diagnosed by our model. This is indicative of two things: (a) the Southern and West-Central states reopened early, which led to a relaxed imposition of quarantine/lockdown measures in these states and consequently a surge in infections was seen, and (b) the North-Eastern states of New York, New Jersey and Illinois reopened late, and even after reopening, a relatively low contact rate was maintained amongst the population, leading to a relatively high magnitude of the imposed quarantine strength, which prevented a surge of infections in these states.

After confirming that our model is able to accurately depict the co-relation between the surge in infections and early reopening in these states through the diagnosed $Q(t)$, we proceed to the second stage of our analysis. In the second stage, we use the diagnosed $Q(t)$ to address the question: How many infected cases would have been reduced, had the worst affected states not reopened early? To answer this question, we simulate the "no-reopening" strategy by assuming that $Q(t)$ is maintained at the value it was before reopening, without decreasing. This simulated $Q(t)$ is shown in Figures 2 b, e, h, k. The flexibility of our model allows us to run our model with this simulated $Q(t)$ for all states considered. To quantify the aleatory uncertainty resulting from random fluctuations in the model, we utilized the chemical Langevin equation extension to the QSIR model whose definition and justification is described in the Methods and Supplemental Information section. This allows us to estimate bootstrapped confidence intervals resulting from 1000 simulations of such a stochastic model, and thus quantify the effect of such a "no-reopening policy" on the epidemic spread. The infected count evolution for the simulated $Q(t)$ without reopening is shown in Figures 2 c, f, i, l (5% and 95% quantiles are shown). We can see that, for all these states, instead of seeing a spike in infections, we would have seen a plateau in the infected case count evolution. The number

Table 1: Drop in quarantine strength function, $Q(t)$ after reopening as discovered by our trained model. $Q(t)$ trained by our model shows a significant drop for all Southern and West-Central states which showed a surge in cases from reopening; whereas the North-Eastern states which showed no surge don't see a drop in $Q(t)$.

State	Reopening date	% increase in daily cases since reopening	Maximum % decrease in $Q(t)$ after reopening
1. Arizona	May 15	+858	+22
2. Florida	May 4	+1393	+10
3. Louisiana	May 15	+193	+30
4. Nevada	May 9	+697	+25
5. Oklahoma	April 24	+477	+29
6. South Carolina	May 4	+999	+71
7. Tennessee	April 30	+279	+44
8. Texas	May 1	+680	+29
9. Utah	May 1	+287	+39
10. New York	May 29	−52	−45
11. New Jersey	June 9	−32	−60
12. Illinois	May 29	−54	−8

Table 2: Infected count reduction by 14 July, 2020, if states had not reopened early, as estimated by our model.

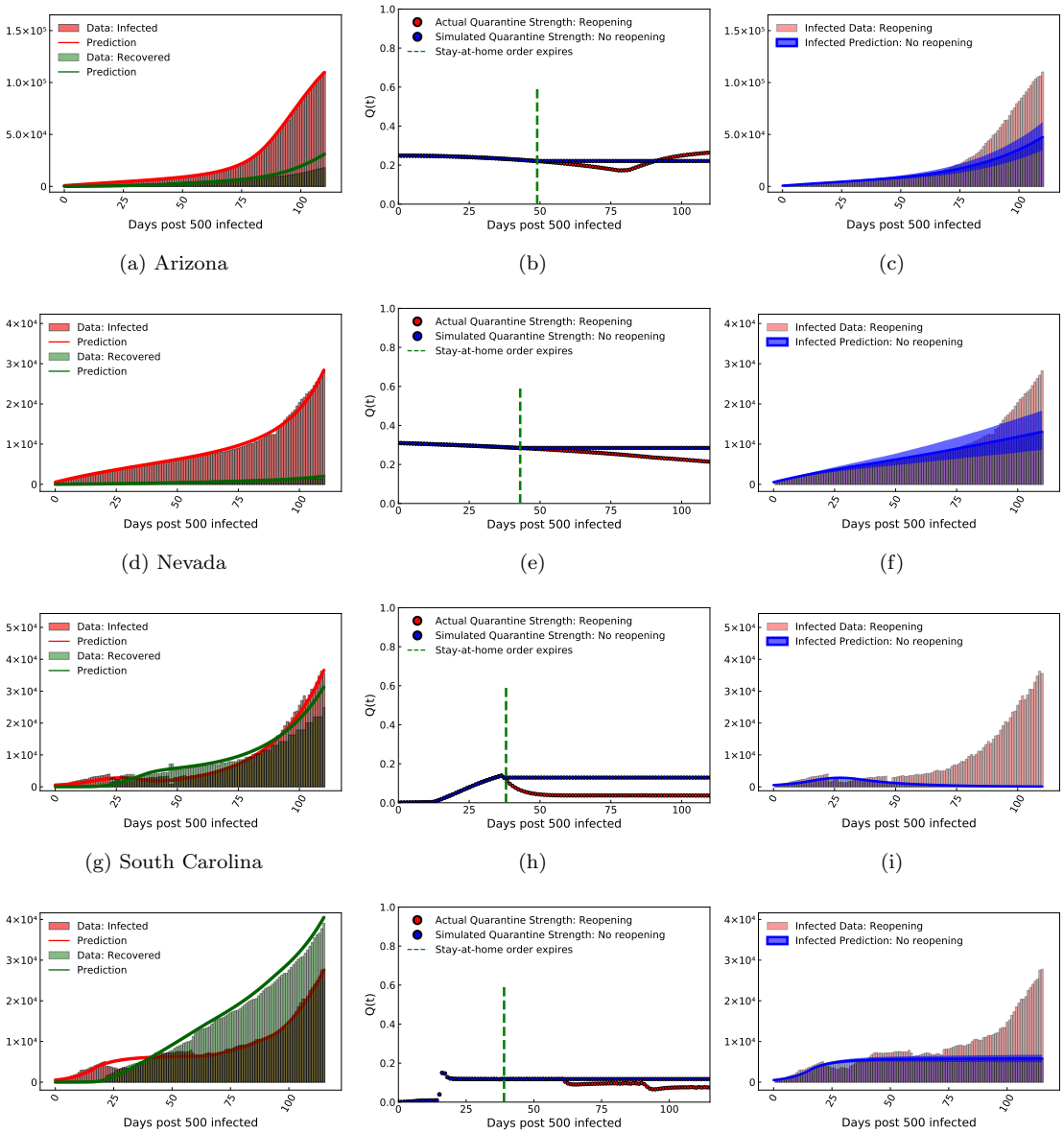
State	% decrease (5% - 95% quantiles)	Mean % decrease	Case reduction	Mean case case reduction
1. Arizona	35 – 62	49	44000 – 79000	63000
2. Florida	20 – 75	49	57000 – 218000	144000
3. Louisiana	37 – 50	44	31000 – 41000	36000
4. Nevada	32 – 68	51	10000 – 20000	15000
5. Oklahoma	46 – 69	58	10000 – 15000	13000
6. South Carolina	83 – 86	84	50000 – 52000	51000
7. Tennessee	41 – 53	47	27000 – 36000	31000
8. Texas	41 – 51	46	115000 – 143000	129000
9. Utah	35 – 47	41	11000 – 14000	12000

and the percentage of infected cases that would have been prevented by July 14 had these states not reopened are shown in Table 2. It is evident that the number of infections could have been reduced by more than 40% in all states considered, with the actual number of infections reduced being more than $100,000$ for the states of Florida and Texas. Even the less populated states of Louisiana, South Carolina and Tennessee show mean infected case reduction values of $44\%, 84\%$ and 47% respectively, which correspond to $36,000, 51,000$, and $31,000$ infected cases reduced.

4 DISCUSSION

In this study, we have developed a novel methodology to quantify the effect of early reopening on the infected case count surge seen during the period of June-July 2020. We have proposed a machine learning model, called the QSIR model, rooted firmly in fundamental epidemiology principles which has the following attributes: (a) it is highly interpretable with few free parameters rooted in an epidemiological model, (b) it relies on only Covid-19 data and not on previous epidemics and (c) it can decompose the infected timeseries data to reveal the quarantine strength/policy variation, $Q(t)$, in the region under consideration. To demonstrate the validity of our model in capturing the actual quarantine policy evolution in a particular region, the model has been applied to 70 countries globally. The quarantine strength behaviour learnt from the model accurately mimics the on-the-ground situation in majorly affected European, South American and Asian continents. The results for this global analysis are hosted at covid19ml.org.[12]

After confirming our belief in the model through a global analysis, we apply the model to



(a) Arizona (b) (c)

(d) Nevada (e) (f)

(g) South Carolina (h) (i)

(j) Tennessee (k) (l)

Figure 2: For the states of Arizona, Nevada, South Carolina and Tennessee, figure shows: (a, d, g, j) Model recovery of infected and recovered case count as of 14 July, 2020. (b, e, h, k) Quarantine strength function as discovered by our trained model (with reopening). This is shown along with the quarantine strength function which we use to simulate strict quarantine without reopening after stay-at-home order was imposed. (c, f, i, l) Estimated infected count if strict quarantine and lockdown measures were followed without reopening (5% and 95% quantiles are shown) as compared to the values corresponding to the actual early reopening scenario.

the Southern and West-Central US states which have shown a massive surge in Covid-19 infected cases since June 2020. We demonstrate that the $Q(t)$ extracted by our model shows a significant drop in value for the Southern and West-Central states which reopened early and showed a surge in infections. The time at which $Q(t)$ starts to decline generally agrees well with the reopening date for the states considered. Since the decline in $Q(t)$ is strongly co-related to the surge of infections and also the reopening date for states which reopened early, we can then simulate the effect of "no-reopening" by maintaining the $Q(t)$ at a constant level after reopening, instead of declining. We show that maintaining a steady imposition of quarantine/lockdown control would have played a massive role in bringing down the infected count by more than 40% in all states considered, with the infections reduced reaching more than $100,000$ for the states of Florida and Texas.

As we continue our fight against Covid-19, our proposed model can be used as a valuable asset to simulate the effect of several reopening strategies on the infected count evolution; for any region under consideration.This makes the proposed model a valuable tool for policy makers and researchers alike. The results of our model should be taken in the context of its assumptions. The model is based on the SIR framework, which assumes a constant, age-independent contact and recovery rate between the infected and susceptible populations. Additionally, we do not consider the spatial heterogeneity in the infected count within a particular state and assume the governing dynamics to be only time-dependent. Consideration of these second-order aspects would further refine the model and would be the subject of future studies.

Our findings highlight that as we continue the fight against Covid-19, it is imperative to reduce the contact between susceptible and infected individuals in public places by formulating robust safety guidelines. Such guidelines implemented and maintained in the affected states would ensure a high level of quarantine strength associated with that state and can prevent a future surge or wave in the Covid-19 infected count timeseries.

5 EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Raj Dandekar, MIT. Email: rajd@mit.edu.

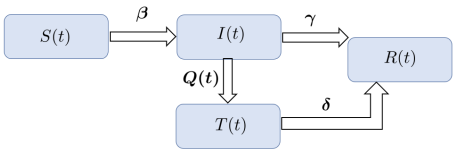
Data and Code Availability

Data for the infected and recovered case count in all regions was obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. All code files and results are publicly available at `https://github.com/RajDandekar/Reopening_ImpactSimulator_US_States`.

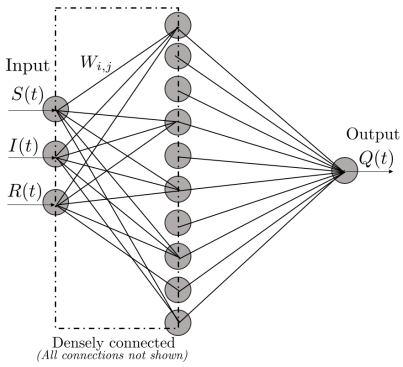
5.1 QSIR Model

In general, neural networks with arbitrary activation functions are universal approximators.[14–16] Unbounded activation functions in particular, such as the rectified linear unit (ReLU) have been known to be effective in approximating nonlinear functions with a finite set of parameters.[17–19] Thus, a neural network solution is attractive to approximate quarantine effects in combination with analytical epidemiological models. The downside is that the internal workings of a neural network are difficult to interpret. The recently emerging field of Scientific Machine Learning[20] exploits conservation principles within a universal differential equation,[21] SIR in our case, to mitigate overfitting and other related machine learning risks.

In the present work, the neural network is trained from publicly available infection and population data for Covid-19 for each state under study.



(a)



(b)

Figure 3: (a) Schematic of the augmented QSIR model considered in the present study. (b) Schematic of the neural network architecture used to learn the quarantine strength function $Q(t)$.

Standard SIR model

The classic SIR epidemiological model is a standard tool for basic analysis concerning the outbreak of epidemics. In this model, the entire population is divided into three sub-populations: susceptible S, infected I and recovered R. The sub-populations' evolution is governed by the following system of three coupled nonlinear ordinary differential equations

$$\frac{dS}{dt} = -\frac{\beta\, S(t)\, I(t)}{N} \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta\, S(t)\, I(t)}{N} - \gamma I(t) \tag{2}$$

$$\frac{dR}{dt} = \gamma I(t). \tag{3}$$

Here, β is the infection rate and γ is the recovery rate, and they are assumed to be constant in time. The total population $N = S(t) + I(t) + R(t)$ is seen to remain constant as well; that is, births and deaths are neglected. The recovered population is to be interpreted as those who can no longer infect others, so it also includes individuals who are deceased due to the infection. The possibility of recovered individuals to become reinfected is accounted for by SEIS models,[22] but we do not use this model here, as the reinfection rate for Covid-19 survivors is considered to be negligible as of now. An important assumption of the SIR models is homogeneous mixing among the subpopulations. Therefore, this model cannot account for social distancing or social network effects. Additionally the model assumes uniform susceptibility and disease progress for every individual, and that no spreading occurs through animals or other non-human means. Alternatively, the SIR model may be interpreted as quantifying the statistical expectations on the respective mean populations, while deviations from the model's assumptions contribute to statistical fluctuations around the mean.

QSIR model: ODE formulation

To study the effect of quarantine control, we start with the SIR epidemiological model. Figure 3a shows the schematic of the modified SIR model, the QSIR model, which we consider. We augment the SIR model by introducing a time varying quarantine strength rate term $Q(t)$ and a quarantined population $T(t)$, which is prevented from having any further contact with the susceptible population. Thus, the term $I(t)$ denotes the infected population still having contact

with the susceptibles, as done in the standard SIR model, while the term $T(t)$ denotes the infected population who are effectively quarantined and isolated. Thus, we can write an expression for the quarantined infected population $T(t)$ as

$$\frac{dT(t)}{dt} = Q(t)I(t) - \delta T(t) \tag{4}$$

Since $Q(t)$ does not follow from first principles and is highly dependent on local quarantine policies, we devised a neural network-based approach to approximate it.
Recently, it has been shown that neural networks can be used as function approximators to recover unknown constitutive relationships in a system of coupled ordinary differential equations.[21,23] Following this principle, we represent $Q(t)$ as an n layer-deep neural network with weights $W_1, W_2 \dots W_n$, activation function r and the input vector $U = (S(t), I(t), R(t))$ as

$$Q(t) = r\left(W_n r\left(W_{n-1} \dots r\left(W_1 U\right)\right)\right) \equiv \mathrm{NN}(W, U) \tag{5}$$

For the implementation, we choose a $n = 2$-layer densely connected neural network with 10 units in the hidden layer and the leaky ReLU activation function. This choice was because we found sigmoidal activation functions to stagnate. The final model was described by 54 tunable parameters. The neural network architecture schematic is shown in figure 3b. The governing coupled ordinary differential equations for the QSIR model are

$$\frac{dS}{dt} = -\frac{\beta\, S(t)\, I(t)}{N} \tag{6}$$

$$\begin{aligned}\frac{dI}{dt} &= \frac{\beta\, S(t)\, I(t)}{N} - (\gamma + Q(t))\, I(t) \\ &= \frac{\beta\, S(t)\, I(t)}{N} - (\gamma + \mathrm{NN}(W, U))\, I(t)\end{aligned} \tag{7}$$

$$\frac{dR}{dt} = \gamma I(t) + \delta T(t) \tag{8}$$

$$\frac{dT}{dt} = Q(t)\, I(t) - \delta T(t) = \mathrm{NN}(W, U)\, I(t) - \delta T(t). \tag{9}$$

Augmented QSIR Model: Initial Conditions
The starting point $t = 0$ for each simulation was the day at which 500 infected cases was crossed, *i.e.* $I_0 \approx 500$. The number of susceptible individuals was assumed to be equal to the population of the considered region. Also, in all simulations, the number of recovered individuals was initialized from data at $t = 0$ as defined above. The quarantined population $T(t)$ is initialized to a small number $T(t = 0) \approx 10$.

Augmented QSIR Model: Parameter estimation
The time-resolved data for the infected, I_{data}, and recovered, R_{data}, for each locale considered is obtained from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The neural network-augmented SIR ODE system was trained by minimizing the mean square error loss function

$$\begin{aligned}L_{\mathrm{NN}}(W, \beta, \gamma, \delta) = \\ &\|\log(I(t) + T(t)) - \log(I_{\mathrm{data}}(t))\|^2 \\ &+ \|\log(R(t)) - \log(R_{\mathrm{data}}(t))\|^2\end{aligned} \tag{10}$$

that includes the neural network's weights W. For most of the states under consideration, W, β, γ, δ were optimized by minimizing the loss function given in (10). Minimization was employed using local adjoint sensitivity analysis[23,24] following a similar procedure outlined in a recent study[21] with the ADAM optimizer[25] with a learning rate of 0.01. $60,000$ iterations were employed for each considered state. For states with a low recovered count: Arizona, Florida, Nevada and Texas, we employed a two stage optimization procedure to find the optimal W, β, γ, δ. In the first stage, (10) was minimized. For the second stage, we fix the optimal γ, δ found in the first stage to optimize

for the remaining parameters: W, β based on the loss function defined just on the infected count as $L(W, \beta) = \|\log(I(t) + T(t)) - \log(I_{\text{data}}(t))\|^2$. In the second stage, we don't include the recovered count $R(t)$ in the loss function, since $R(t)$ depends on γ, δ which have already been optimized in the first stage. By placing more emphasis on minimizing the infected count, such a two stage procedure leads to much more accurate model estimates when the recovered data count is low. The iterations required for convergence in both stages varied based on the region considered and generally ranged from $40000 - 60000$.
In all states considered in the present study, we trained the model using data starting from the dates when the 500th infection was recorded in each region and up to July 14, 2020. For each state considered, $Q(t)$ denotes the rate at which infected persons are effectively quarantined and isolated from the remaining population, and thus gives composite information about (a) the effective testing rate of the infected population as the disease progresses and (b) the intensity of the enforced quarantine as a function of time.
This QSIR ODE framework applied on the infected and recovered data is used to estimate the quarantine strength function Q(t) in a particular state as shown in the first and second columns of figure 2.

QSIR Model: SDE formulation

The ODE modelling framework described above is a deterministic approach to model transfer of species (here: people) from one compartment to another through different reaction channels. Such a deterministic approach ignores any random fluctuations during species transfer from one compartment to the other. To include such stochastic effects and thus get a measure of the model uncertainty, we note that the augmented SIR framework derives from the chemical master equation which descibes the time evolution of the probability of such a system of interacting species to be in a given state at a given time (details in Supplementary Information). Although the chemical master equation cannot be solved analytically, under certain conditions, it can be distilled down to a stochastic differential equation (SDE) which captures the fluctuations in species transfer as random walks. Such an SDE, also known as the Chemical Langevin Equation, is thus based on the underlying ODE framework (macroscopic picture) and also includes stochastic effects reminiscent of microscopic modelling. In fact, in the Supplementary Information, we show that the microscopic simulation, macroscopic ODE formulation and the Chemical Langevin Equation (which acts as a bridge between the two) are all equivalent to each other.

The equivalent stochastic formulation or the Chemical Langevin equation for the augmented SIR model is

$$dS = -\left[\frac{\beta\, S(t)\, I(t)}{N}\right] dt - \sqrt{\left[\frac{\beta\, S(t)\, I(t)}{N}\right]} dW_1(t) \tag{11}$$

$$dI = \left[\frac{\beta\, S(t)\, I(t)}{N} - \gamma I(t) - Q(t) I(t)\right] dt + \sqrt{\frac{\beta\, S(t)\, I(t)}{N}} dW_1(t) - \sqrt{\gamma I(t)} dW_2(t) - \sqrt{Q(t) I(t)} dW_3(t) \tag{12}$$

$$dR = [\gamma I(t) + \delta T(t)]\, dt + \sqrt{\gamma I(t)} dW_2(t) + \sqrt{\delta T(t)} dW_4(t) \tag{13}$$

$$dT = [Q(t)\, I(t) - \delta T(t)]\, dt + \sqrt{Q(t) I(t)} dW_3(t) - \sqrt{\delta T(t)} dW_4(t) \tag{14}$$

In (11), $W_i(t) \sim N(0, t)$ is a normally distributed random variable with mean zero and variance t or $dW_i(t) \sim N(0, dt)$. It should also be noted that each $W_i(t)$ represents an independent Brownian motion. The simulations were performed using the Catalyst.jl software in Julia using the LambaEM algorithm based on.[26] 1000 trajectories were simulated for each state.

This QSIR SDE framework along with the simulated quarantine functions for no reopening is used to predict the new infected case count and hence estimate the reduction in the number of infected cases under the simulated no-reopening quarantine function. The results are shown as 5% and 95% quantiles in the third column of figure 2.

Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is defined as

$$\text{MAPE} = \frac{100}{N} * \sum \frac{[I(t) + T(t) + R(t)] - [I_{\text{data}}(t) + R_{\text{data}}(t)]}{[I_{\text{data}}(t) + R_{\text{data}}(t)]} \tag{15}$$

where N is the number of observations.

6 Supplementary Information

Model-diagnosed quarantine strength for North-Eastern US states

Figure 4 shows the application of the model to the north-eastern states of New York, New Jersey and Illinois along with the diagnosed quarantine strength function $Q(t)$ for these states. These states do not show a decline in $Q(t)$. This corresponds well to the delayed reopening and generally stronger quarantine measures employed in the North-Eastern US states. Since $Q(t)$ does not decrease, these states did not show a surge in infections starting June 2020, unlike their Southern and West-Central counterparts. The difference in these results between the North-Eastern and Southern, West-Central states indicates two things: (a) it strengthens the validity of our proposed model in capturing the real-time reopening scenario in different states through the evolution of the diagnosed $Q(t)$, and, more importantly, (b) it further validates the role played by early reopening in reducing $Q(t)$ and subsequently leading to a surge of new infected cases in the Southern and West-Central US states.

Impact of early reopening on the states of Louisiana, Florida, Oklahoma, Texas and Utah

Figure 5, 6 implements a similar analysis to study the effect of early reopening for the states of Louisiana, Nevada, Oklahoma, Texas and Utah, as done for the states of Arizona, Nevada, South Carolina and Tennessee. Similar to the states considered in the main text, we see that all of these states show a decline in $Q(t)$ starting around the time when these states were reopened. If these states were not reopened early, a large number of infections would have been reduced as demonstrated in Table 1 of the main text.

Equivalence between the ODE model and the Chemical Langevin SDE model

This analysis heavily borrows from the pioneering work done by Gillespie.[?] In this section, we will establish that the deterministic ODE model and the stochastic Chemical Langevin equation originate from a common expression: the chemical master equation,[27] and are closely linked to one another. Following is the notation we will use, in accordance with[?] We consider N compartments: $S_1, S_2 \dots S_N$ and R reaction channels: $R_1, R_2 \dots R_M$ in a fixed volume Ω. In our case, we have $N = 4$ (S, I, R, T) compartments and $R = 4$ reaction channels. We denote the dynamical state of the system at any time t as $X(t) = (X_1(t), X_2(t) \dots X_N(t))$ where

- $X_i(t)$: total number of S_i molecules (in our case: individuals) in the system.
- Propensity function $a_j(x)dt$: probability that a reaction R_j will occur somewhere in Ω in the next time interval [t, t+dt] for $j = 1, 2 \dots M$.
- State change vector ν_j whose ith component is defined by $\nu_{j,i}$: change in the number of S_i molecules produced by one R_j reaction for $i = 1, 2 \dots N$, $j = 1, 2 \dots M$. In our case $\nu_{j,i} = \pm 1$.

From the definition of $a_j(x)dt$, we can write the probability of the system being in state x at time $t + dt$ (we take the sum of all mutually exclusive ways either through one reaction or no reaction in [t, t+dt]):

$$P(x,t+dt|x_0,t_0) = P(x,t|x_0,t_0)\left[1-\sum_{j=1}^{M} a_j(x)dt\right] + \sum_{j=1}^{M}\left[P(x-\nu_j,t|x_0,t_0)a_j(x-\nu_j)dt\right], \tag{16}$$

Taking the limit of (16) as dt -¿ 0 leads to the **chemical master equation**

$$\frac{\partial P(x,t|x_0,t_0)}{\partial t} = \sum_{j=1}^{M}\left[a_j(x-\nu_j)P(x-\nu_j,t|x_0,t_0) - a_j(x)P(x,t|x_0,t_0)\right] \tag{17}$$

.

Macroscopic picture: Deterministic model relation to the chemical master equation:

Multiplying the chemical master equation (17) by x_i and summing over all x, we obtain for the mean of $X_i(t)$

$$\frac{d\langle X_i(t)\rangle}{dt} = \sum_{j=1}^{M} \nu_{ji}\langle a_j(X(t))\rangle \qquad (i = 1,2\ldots N) \tag{18}$$

Thus, whenever fluctuations are not important, the species populations evolve deterministically according to the following set of ordinary differential equations

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^{M} \nu_{ji} a_j(X(t)) \qquad (i = 1,2\ldots N) \tag{19}$$

(19) is the basis for the classical SIR epidemiological equations, and we see how they evolve from the chemical master equation (17).

$$\frac{dS}{dt} = -\frac{\beta\, S(t)\, I(t)}{N} \tag{20}$$

$$\frac{dI}{dt} = \frac{\beta\, S(t)\, I(t)}{N} - (\gamma + Q(t))\, I(t)$$

$$= \frac{\beta\, S(t)\, I(t)}{N} - (\gamma + \mathrm{NN}(W,U))\, I(t) \tag{21}$$

$$\frac{dR}{dt} = \gamma I(t) + \delta T(t) \tag{22}$$

$$\frac{dT}{dt} = Q(t)\, I(t) - \delta T(t) = \mathrm{NN}(W,U)\, I(t) - \delta T(t). \tag{23}$$

The ODE system used in the present study shown in (5-8), is of the form (19).

Microscopic picture: Stochastic Simulation Algorithm and its relation to the master equation:

Another consequence of the master equation (17) is the existence and form of the next-reaction density function $p(\tau, j|x,t)$, which is defined as

- $p(\tau, j|x,t)d\tau =$ probability that given $X(t) = x$, the next reaction in Ω will occur in $[t+\tau, t+\tau+d\tau]$, and will be an R_j reaction

Since $\sum_j a_j(x)dt$ is the probability that some reaction occurs in the time interval dt, the probability that a time interval τ is spent without any reaction occuring is given by the exponential distribution: $\mathrm{Exp}(\sum_j a_j(x)\tau)$. Thus, we obtain for $p(\tau, j|x,t)$

$$p(\tau, j|x,t) = a_j(x)\mathrm{Exp}\left(\sum_{k=1}^{M} a_k(x)\tau\right) \qquad (0 \le \tau < \infty; j = 1,2\ldots M) \tag{24}$$

(24) is the basis for the stochastic simulation algorithm in which Monte-Carlo techniques are used to construct unbiased realizations of the process $X(t)$. A typical algorithm for stochastic simulation of this kind, is the Gillespie Algorithm[?] which can be viewed as a discrete space continuous time Markov jump process, with exponentially distributed jump times.

Chemical Langevin Equation: Bridging the gap between macroscopic and microscopic models:

Let the state of the system $X(t)$ at the current time t be x_t. Let $K_j(x_t, \tau)$ be the number of R_j reactions that occur in the time interval [t, t+dt]. Thus, the number of S_i molecules in the system at time $t + \tau$ will be

$$X_i(t+\tau) = x_{ti} + \sum_{j=1}^{M} K_j(x_t, \tau)\nu_{ji} \qquad (i = 1, 2 \ldots N) \tag{25}$$

? approximated K_j by imposing the following conditions

- **Condition 1: No propensity function change** This condition requires τ to be small enough so that none of the propensity functions $a_j(x)$ change noticeably. The propensity functions then satisfy

$$a_j(X(t')) \approx a_j(x_t) \qquad \forall t \in [t, t+\tau], \forall j \in [1, M] \tag{26}$$

  Due to this condition, $K_j(x_t, \tau)$ will be a statistically independent Poisson random variable $P_j(a_j(x_t), \tau)$. Thus (25) simplifies to

$$X_i(t+\tau) = x_{ti} + \sum_{j=1}^{M} \nu_{ji} P_j(a_j(x_t), \tau) \qquad (i = 1, 2 \ldots N) \tag{27}$$

- **Condition 2: Large number of reaction occurrences:** This condition requires τ to be large enough so that the expected number of occurrences of each reaction channel R_j in $[t, t+\tau]$ is much larger than 1. Thus

$$\langle P_j(a_j(x_t), \tau)\rangle = a_j(x_t)\tau \gg 1, \qquad \forall j \in [1.M]. \tag{28}$$

  This condition enables us to approximate each Poisson variable $P_j(a_j(x_t), \tau)$ by a normal random variable with the same mean and variance.

  Thus, (27) further simplifies to

$$X_i(t+\tau) = x_{ti} + \sum_{j=1}^{M} \nu_{ji} N_j(a_j(x_t)\tau, a_j(x_t)\tau) \qquad (i = 1, 2 \ldots N) \tag{29}$$

  where $N(m, \sigma^2)$ denotes the normal random variable with mean m and variance σ^2. Using $N(m, \sigma^2) = m + \sigma N(0, 1)$, denoting the time interval τ by dt and the unit normal random variable $N_j(0, 1)$ as $N_j(t)$, we obtain

$$X_i(t+dt) = X_i(t) + \sum_{j=1}^{M} \nu_{ji} a_j(X(t))dt + \sum_{j=1}^{M} \nu_{ji} a_j^{1/2}(X(t)) N_j(t)(dt)^{1/2} \qquad (i = 1, 2 \ldots N) \tag{30}$$

  (30) can be written as a stochastic differential equation as

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^{M} \nu_{ji} a_j(X(t)) + \sum_{j=1}^{M} \nu_{ji} a_j^{1/2}(X(t))\Gamma_j(t) \tag{31}$$

  where $\Gamma_j(t)$ are temporally uncorrelated, statistically independent Gaussian white noise processes.

  (31) is the Langevin equation, and it derives from the master equation provided that Condition 1 and Condition 2 are satisfied.

Table 3: Mean Absolute Percentage Error (MAPE) values are shown along with the number of epochs required for and the number of parameters optimized, for all states considered.

State	Model MAPE	Epochs	Parameters optimized
1. Arizona	5.4%	10^5	54
2. Florida	18.7%	10^5	54
3. Louisiana	12%	12^5	54
4. Nevada	3.14%	18^5	54
5. Oklahoma	7.9%	12^5	54
6. South Carolina	11.7%	12^5	54
7. Tennessee	6.9%	12^5	54
8. Texas	10.4%	24^5	54
9. Utah	3.79%	12^5	54

The Langevin equation (31) form of the ODE system (5-8) leads to the stochastic differential equation used in the current study

$$dS = -\left[\frac{\beta\, S(t)\, I(t)}{N}\right] dt - \sqrt{\left[\frac{\beta\, S(t)\, I(t)}{N}\right]} dW_1(t) \tag{32}$$

$$dI = \left[\frac{\beta\, S(t)\, I(t)}{N} - \gamma I(t) - Q(t) I(t)\right] dt + \sqrt{\frac{\beta\, S(t)\, I(t)}{N}} dW_1(t) - \sqrt{\gamma I(t)} dW_2(t) - \sqrt{Q(t) I(t)} dW_3(t) \tag{33}$$

$$dR = [\gamma I(t) + \delta T(t)]\, dt + \sqrt{\gamma I(t)} dW_2(t) + \sqrt{\delta T(t)} dW_4(t) \tag{34}$$

$$dT = [Q(t)\, I(t) - \delta T(t)]\, dt + \sqrt{Q(t) I(t)} dW_3(t) - \sqrt{\delta T(t)} dW_4(t) \tag{35}$$

In (32), $W_i(t) \sim N(0, t)$ is a normally distributed random variable with mean zero and variance t or $dW_i(t) \sim N(0, dt)$. It should also be noted that each $W_i(t)$ represents an independent Brownian motion.

Comparison of the macroscopic, microscopic and Langevin SDE model for our study

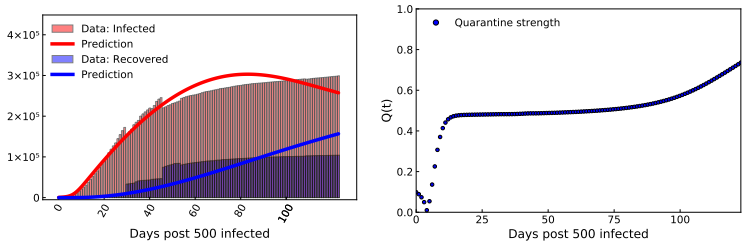
Figure 7a shows that the microscopic Stochastic Simulation Gillespie Algorithm and the ODE model presented in Equation (6-9) in the main text show a good agreement with each other. Figure 7b shows the comparison of the Chemical Langevin SDE model shown in (32) ran for 1000 trajectories and the ODE model; which also show a good agreement. Thus, we have shown the equivalence between the microscopic, macroscopic and the Chemical Langevin model for our study. This equivalence allows us to add fluctuating components to the standard deterministic SIR model as shown in (32) and quantify the uncertainty resulting from these fluctuations.

Model specifications for each state

Table 3 shows the Model Mean Absolute Percentage Error (MAPE), epochs needed for convergence and number of parameters optimized for the different states considered.

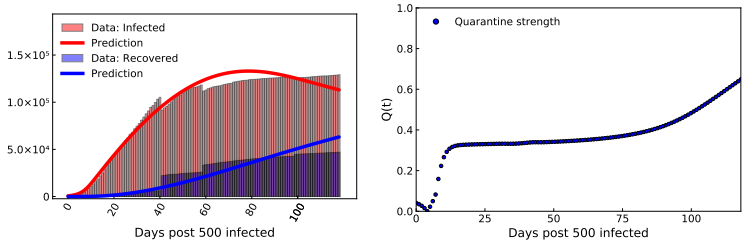
7 ACKNOWLEDGEMENTS

This effort was partially funded by the Intelligence Advanced Reseach Projects Activity (IARPA). We are grateful to Haluk Akay, Hyungseok Kim and Wujie Wang for helpful discussions and suggestions.



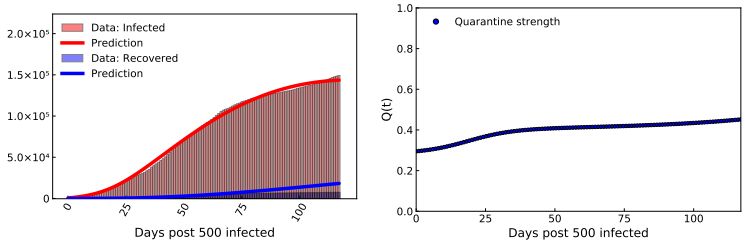
(a) New York

(b)



(c) New Jersey

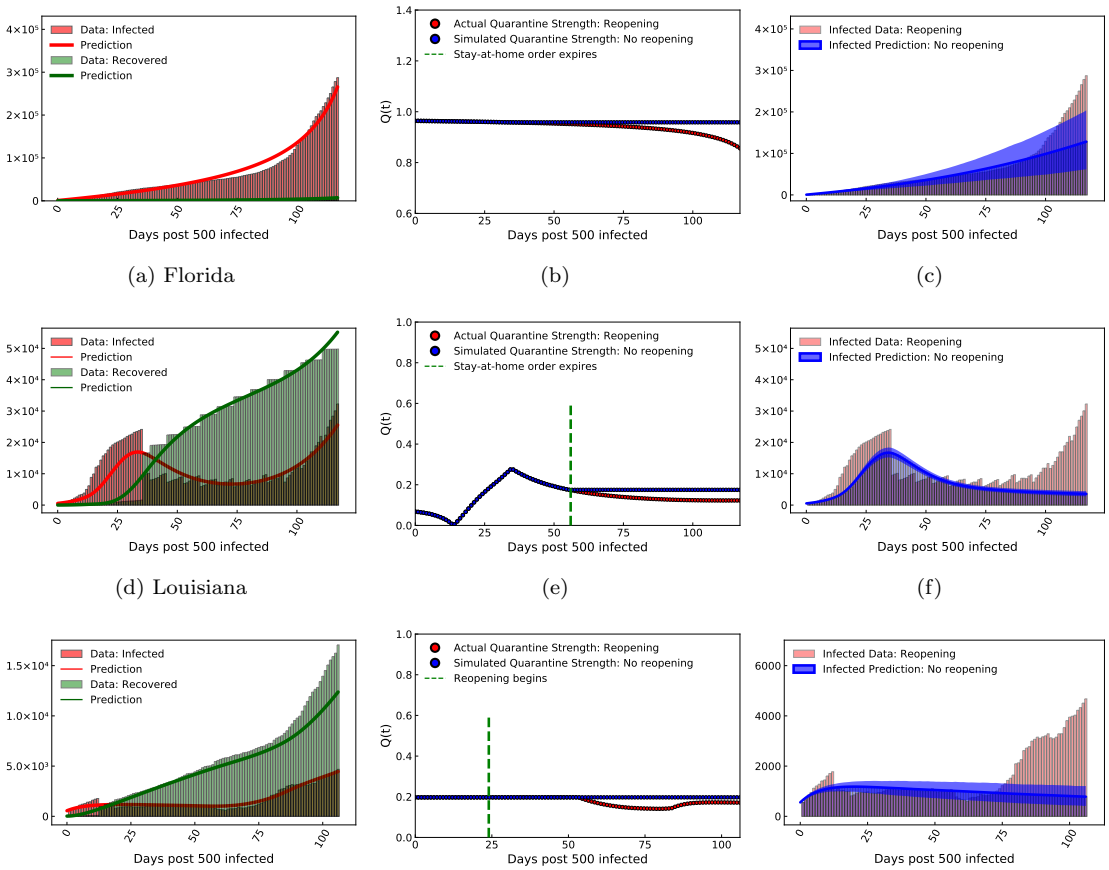
(d)



(e) Illinois

(f)

Figure 4: For the states of New York, New Jersey and Illinois, figure shows: (a, c, e) Model recovery of infected and recovered case count trained until 14 July, 2020. (b, d, f) Quarantine strength function as discovered by our trained model

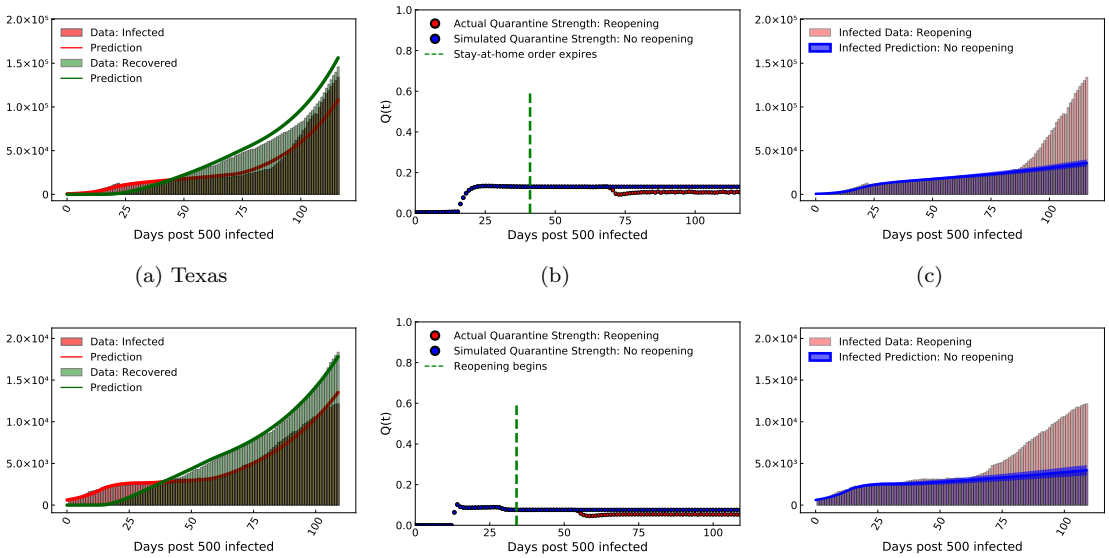


(a) Florida (b) (c)

(d) Louisiana (e) (f)

(g) Oklahoma (h) (i)

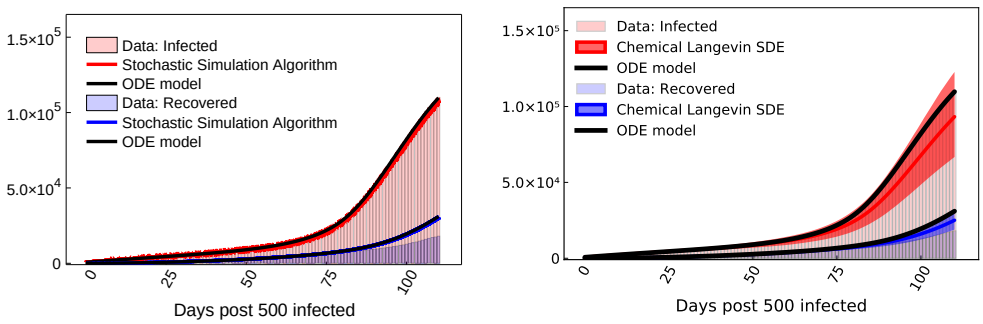
Figure 5: For the states of Louisiana, Nevada and Oklahoma: (a, d, g) Model recovery of infected and recovered case count as of 14 July, 2020. (b, e, h) Quarantine strength function as discovered by our trained model (with reopening). This is shown along with the quarantine strength function which we use to simulate strict quarantine without reopening after stay-at-home order was imposed. (c, f, i) Estimated infected count if strict quarantine and lockdown measures were followed without reopening as compared to the values corresponding to the actual early reopening scenario.



(a) Texas (b) (c)

(d) Utah (e) (f)

Figure 6: For the states of Texas and Utah: (a, d) Model recovery of infected and recovered case count as of 14 July, 2020. (b, e) Quarantine strength function as discovered by our trained model (with reopening). This is shown along with the quarantine strength function which we use to simulate strict quarantine without reopening after stay-at-home order was imposed. (c, f) Estimated infected count if strict quarantine and lockdown measures were followed without reopening as compared to the values corresponding to the actual early reopening scenario.



(a) Arizona (b)

Figure 7: (a) Comparison of the microscopic Stochastic Simulation Gillespie Algorithm and the ODE model presented in Equation (6-9) in the main text. (b) Comparison of the Chemical Langevin SDE model shown in (32) ran for 1000 trajectories ($5\% and 95\%$ quantiles are shown) and the ODE model.

8 AUTHOR CONTRIBUTIONS

R.D. and G.B. designed the research. C.R. and R.D. designed the model framework. R.D. and E.W. applied the model to all the states considered. R.D., C.R., E.W. and G.B. wrote the study.

9 DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

References

[1] Chan, J, Yuan, S, Kok, K, To, K, Chu, H, Yang, J, Xing, F, Liu, J, Yip, C, Poon, R, & Tsoi, H. (2020) *The Lancet* **395**.

[2] CDC. (2020) *CDC Report.*

[3] WHO. (2020) *WHO Report.*

[4] Holshue, M. L, DeBolt, C, Lindquist, S, Lofy, K. H, Wiesman, J, Bruce, H, Spitters, C, Ericson, K, Wilkerson, S, & Tural, A. o. (2020) *New England Journal of Medicine.*

[5] Carey, B & Glanz, J. (2020) *The New York Times.*

[6] Meyer, R & Madrigal, A. C. (2020) *The Atlantic.*

[7] Bellware, K, Hawkins, D, Knowles, H, Denham, H, Kornfield, M, Brice-Saddler, M, Iati, M, & Partlow, J. (2020) *The Washington Post.*

[8] Knowles, H, Wagner, J, Shaban, H, Taylor, A, Copeland, K, Buckner, C, Kornfield, M, & Itkowitz, C. (2020) *The Washington Post.*

[9] Gamio, L. (2020) *The New York Times.*

[10] Lee, J. C, Mervosh, S, Avila, Y, Harvey, B, & Matthew, A. L. (2020) *The New York Times.*

[11] Elassar, A. (2020) *Cable News Network.*

[12] Dandekar, R, Rackauckas, C, & Barbastathis, G. (2020) *(Accepted for publication) Cell Patterns.*

[13] Dandekar, R & Barbastathis, G. (2020) *medRxiv.*

[14] Cybenko, G. (1989) *Math Control, Sig. & Sys.* **2**, 303–314.

[15] Hornik, K. (1991) *Neural Networks* **4**, 251–257.

[16] Sho Sonoda & Noboru Murata. (2017) *Appl. Comp. Harmonic Anal.* **43**, 233–268.

[17] Glorot, X, Bordes, A, & Bengio, Y. (2011) *Deep sparse rectifier neural networks.* pp. 315–323.

[18] Goodfellow, I, Wade-Parsley, D, Mirza, M, Courville, A, & Bengio, Y. (2013) *Maxout networks.* Vol. 28, pp. 1319–1327.

[19] Dahl, G. E, Sainath, T. N, & Hinton, G. E. (2013) *Improving deep neural networks for LVCSR using rectified linear units and dropout.* pp. 8609–8613.

[20] Baker, N, Alexander, F, Bremer, T, Hagberg, A, Kevrekidis, Y, Najm, H, Parashar, M, Patra, A, Sethian, J, Wild, S, et al. (2019).

[21] Rackauckas, C, Ma, Y, Martensen, J, Warner, C, Zubov, K, Supekar, R, Skinner, D, & Ramadhan, A. (2020) *arXiv preprint arXiv:2001.04385.*

[22] Mukhopadhyay, B & Bhattacharyya, R. (2008) *Nonlinear Analysis: Real World Applications* **9**, 585–598.

[23] Rackauckas, C, Innes, M, Ma, Y, Bettencourt, J, White, L, & Dixit, V. (2019) *CoRR* **abs/1902.02376**.

[24] Cao, Y, Li, S, Petzold, L, & Serban, R. (2003) *SIAM journal on scientific computing* **24**, 1076–1089.

[25] Kingma, D & Ba, J. (2014) *arXiv preprint arXiv:1412.6980.*

[26] Rackauckas, C & Nie, Q. (2017) *Discrete and continuous dynamical systems. Series B* **22**, 2731.

[27] McQuarrie, D. A. (1967) *Journal of applied probability* **4**, 413–478.