Characterizing the Divergence Between Two Different Models for Fitting and Forecasting the COVID-19 Pandemic

Tian Gan and Long Ma

May 4, 2021

Noname manuscript No.
(will be inserted by the editor)

Characterizing the divergence between two different models for fitting and forecasting the COVID-19 pandemic

Tian Gan · Long Ma

Received: date / Accepted: date

Abstract Since the novel Coronavirus (COVID-19) has been announced as a global pandemic, researchers from different disciplines have attempted to describe and forecast the spread of COVID-19. Some recent studies try to predict the future trend of the COVID-19 pandemic by deep learning, e.g., the long short-term memory (LSTM), but most works focus on the compartmental epidemic model based curve fitting and forecast. The susceptible-infected-removed (SIR) model and the susceptible-exposed-infected-removed (SEIR) model are two most commonly used compartmental models. The question is to what extent the choice of epidemic models will affect the fitting and long-term forecast performance. In this work, we compared the fitting and prediction performance by considering and ignoring the exposed state to characterize the divergence between these two different models.

Keywords COVID-19 pandemic · Forecast · Exposed state

1 Introduction

It is reported by the World Health Organization (WHO) that the Coronavirus disease 2019 (COVID-19), which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been a global pandemic since early 2020 [1]. Due to the huge impacts of COVID-19 on people's daily life and the economy, many works attempt to find out the spreading mechanism of

T. Gan
Faculty of Electrical Engineering, Mathematics and Computer Science, P.O Box 5031, 2600 GA Delft, The Netherlands;

L. Ma
Faculty of Electrical Engineering, Mathematics and Computer Science, P.O Box 5031, 2600 GA Delft, The Netherlands;
E-mail: l.ma-2@tudelft.nl

COVID-19 and predict how long the pandemic will last and how many people will be finally affected or deceased [2]. One typical way to mathematically describe the dynamics of COVID-19 is to use the compartmental models in epidemiology which assumes that the spreading is in an infinite well-mixed population. The Susceptible-Infected-Removed (SIR) model is one of the most basic compartmental models to describe the COVID-19 pandemic. In the SIR model [3,4,5,6], there are three compartments: the fraction of susceptible individuals (S), the fraction of active infectious individuals (I) and the fraction of the cumulative removed individuals (R). Another common used compartmental epidemic model is the Susceptible-Exposed-Infected-Removed (SEIR) model that considered an incubation period during which the people carried the virus but still cannot infect the susceptible individuals [7,8,9,10,11,12]. There are many variations on the SIR and SEIR models, such as considering the unreported individuals [13], population-level [14] and the underlying network structure [3,15,16].

Although it has been proved that there are on average 5 to 6 days incubation period [17,18] for individuals infected with SARS-CoV-2, the exposed state is still not considered in many forecast-related studies [9,10,3,4,19]. An important question is whether the fitting and forecast accuracy will be significantly affected if the exposed state is not considered. It would be difficult to figure out this question by directly working on the real data since there are many external factors influencing the outcome apart from the basic parameters in the classic SIR or SEIR model. For example, it has been proved that the infection rate is changing with time because of the weather [20]. Moreover, many policies, e.g., quarantine and wearing masks, may also significantly affect the spreading process. We thus first compare the fitting and forecast accuracy on the synthetic data. Our results reveal that the fitting performance will be similar no matter the exposed state is considered or not. However, the results of long-term forecast could be significantly different by considering or neglecting the exposed state. We finally fit and forecast the real data for four different countries[1] by the SIT and SEIT models to back up our conclusions.

2 Compartmental models to describe the COVID-19 pandemic

The classic SIR and SEIR models have been implemented in plenty of prior studies to describe the epidemic outbreak [21,22]. In the SIR model, the susceptible individuals are infected by the infectious individuals with an infection rate β and the infectious individuals are removed (recovered or deceased) with a removed rate δ. In the SEIR model, there is an exposed state E between the susceptible state S and the infectious state I. The susceptible individuals are infected by the infectious individuals and turn to be the exposed individual with an infection rate β. The exposed individual turn to be infectious with a rate σ, where σ^{-1} denotes the average incubation period. There is a common

[1] https://covid19.who.int/info/

misconception that the tested data is equal to the infected data. In reality, the tested infectious individuals are almost impossible to infect the susceptible individuals because the infected people will be quarantined once tested positive. We thus revise the SIR model and SEIR model as the Susceptible-Infected-Tested (SIT) model and Susceptible-Exposed-Infected-Tested (SEIT) model respectively. Specifically, the removed state R is revised to be the tested state T and the removed rate δ is revised to be the tested rate γ. Then the revised differential equations are

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dT}{dt} &= \gamma I \end{aligned} \tag{1}$$

for the SIT model and

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dE}{dt} &= \frac{\beta SI}{N} - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dT}{dt} &= \gamma I \end{aligned} \tag{2}$$

for the SEIT model.

According to the recent researches [18] and [23], the incubation period of COVID-19 is usually between 2 and 14 days. Furthermore, the research [23] shows that the estimated mean incubation period is 5.5 days. Thus, we take the latency period as 5.5 days in our work. At each day k, the fraction of daily new tested individuals is $T_{daily}[k] = T[k] - T[k-1]$. In real data, the time series cannot be as smooth as the curves in the SEIT model and the noise increases with the number of daily tested individuals. We thus generate the daily tested data with noise at each day $T_{noise}[k]$ by the normal distribution with the mean equals $T_{daily}[k]$ and the standard deviation equals $10T_{daily}[k]$. In this work, we set the parameters $\beta = 0.5$ and $\gamma = 0.3$. The initial fractions of S, E, I, T for the SEIT model are set to be $[0.999, 0.001, 0, 0]$, respectively. Figure 2 shows the curves before and after adding the Gaussian noise. We set the curves after adding the Gaussian noise as the benchmark to be fitted and forecasted. **What we will do on the synthetic data is to fit the curves in the first 30 days and forecast prevalence in the last 30 days.** The fitting error and forecast error can be measured by the Root Mean Square Error (RMSE) [24]:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{L}(\widetilde{T}[k] - T[k])}{L}}, \tag{3}$$

where L denotes the length of time series, $T[k]$ denotes the benchmark data and $\widetilde{T}[k]$ is the fit or forecast result.

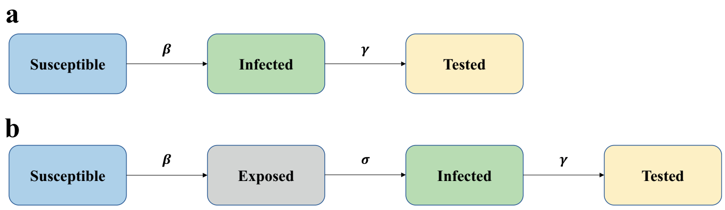

Fig. 1 Diagram of the Susceptible-Infected-Tested (SIT) model (figure a) and the Susceptible-Exposed-Infected-Tested (SEIT) model (figure b) that applied to describe the COVID-19 pandemic.

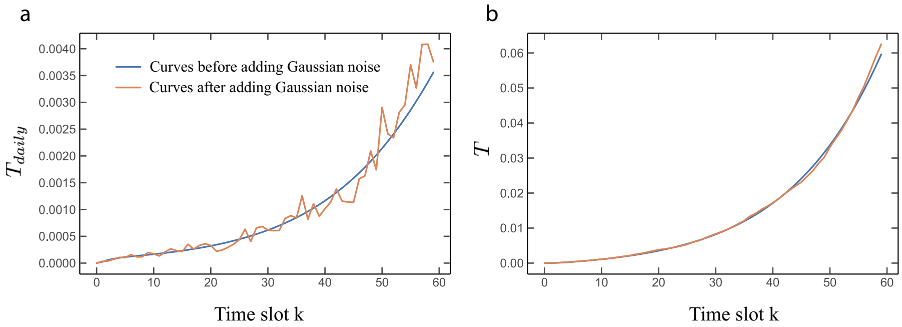

Fig. 2 Curves before and after adding the Gaussian noise. The left figure is for the fraction of daily tested individuals and the right figure is for the fraction of cumulative tested individuals.

3 Results and discussion

We first investigate how the fit or forecast error will vary with the epidemic parameters β and γ. We go over all pairs of the infection rate β and the tested rate γ in the range of $(0, 1)$ with the interval of 0.01. Based on each parameter pairs β and γ, we can generate the curves $\widetilde{T}$ by numerically solving Equations (2) and further measure the difference between $\widetilde{T}$ and the benchmark curve T by RMSE as shown in Equation (3). In order to illustrate the RMSE value clearer, we use the $log(1/RMSE)$ since this value is more suitable for showing small values. Figure 3.c shows the heatmap of $log(1/RMSE)$ with different parameter pairs β and γ. In the heatmaps, we map the RMSE values to $log(1/RMSE)$ and the fitting problem corresponds to the maximization problem $w.r.t$ values in heatmap. It reveals that there are many parameter pairs β and γ that can fit well with the benchmark curve. We further try to fit the benchmark curve (generated by the SEIT model) by the SIT model as shown in Equation (1). The heatmap of $log(1/RMSE)$ with different parameter pairs β and γ is shown in Figure 3.d. It reveals that the SIT model can also fit the benchmark curve generated by the SEIT model well, but the fitted parameters are all far from the real parameters β and γ.

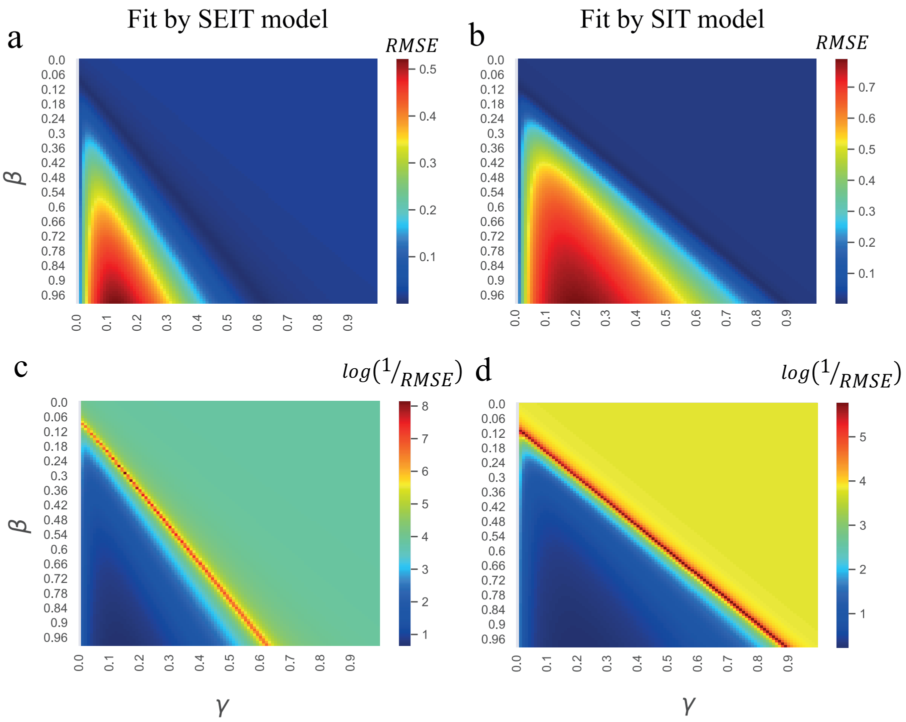

Fig. 3 Heatmaps on the RMSE between the benchmark curve and the estimate curve (generated by the SEIT model or SIT model) with both β and γ in range of $(0, 1)$ with the interval of 0.01. The benchmark curve is generated by the SEIT model with Gaussian noise. The upper figures are for the RMSE values and the bottom figures are for the $log(1/RMSE)$ values, which can find the positions of best fit easier. There are many different parameters in both SIT model and SEIT model whose curves are very similar to the benchmark curve. It also means that, given a curve generated by the SEIT model, it is likely to find a similar curve that is generated by the SIT model. An inference is, if one fits the real COVID-19 data by SIT or SEIT model, the results can be very similar.

After compared the fitting performance between the SIT and SEIT models, we further investigate the forecast performance. Here we propose a hill-climbing algorithm to fit the curves. The details of this algorithm are shown in Appendix A. By fitting the curves with different epidemic models, we can respectively derive the estimated parameters β, γ and R_0. Note that the basic reproduction number $R_0 = \beta/\gamma$. Figure 4 shows the estimated parameters by fitting the benchmark curve based on the SIT and SEIT models. It indicates that the inferred parameters achieved from the SEIT model are closer to the true values than the results from the SIT model. We future predict the last 30 days by generating curves using the inferred parameters β and γ. The left figure of Fig. 5 shows that the predicted curves fitted by the SEIT model are closer to the benchmark curves. We also calculate the prediction error for each day of the 30-day prediction results, which is shown in the right figure of Fig. 5. During the prediction period, the relative errors of the prediction results derived from the SIT and SEIT models are gradually increasing and the prediction error of each day from the SEIT model is always smaller than the results from SIT model.

Finally, we fit and forecast the real data from various countries by SIT and SEIT models as shown in Fig. 6. For the fitting results, we can see that SIT and SEIT model have a similar performance. For the prediction results, the curves from France, Belgium and Italy fitted by SEIT model have a lower RMSE value corresponding to the curves fitted by SIT model. But for the real data from India, the forecast by SIT model has a higher accuracy.

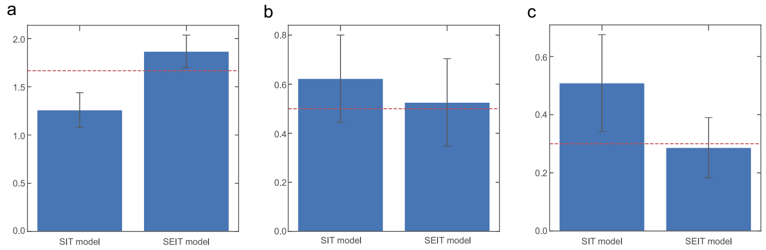

Fig. 4 Output of the estimated parameters R_0 (figure a), β (figure b) and γ (figure c) from the SEIT model and the SIT model. The red lines mark the true values. Although the fitting results from the SIT and SEIT model are similar, the estimated parameters are far different for each model. The errorbar denotes the standard deviation.

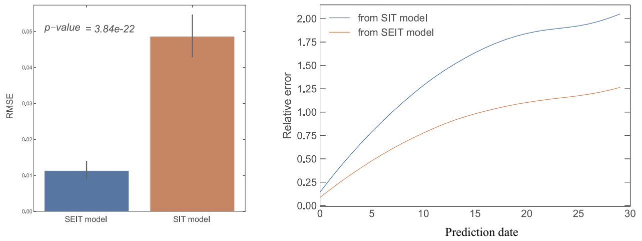

Fig. 5 Forecast results by the SEIT and SIT model. (left) The RMSE between the fit curves and benchmark curves by SEIT model and SIT model. The errorbar denotes the standard deviation. (right) Relative errors of the forecast results in each day. Both predictions are derived by using the SIT and SEIT model separately. It reveals that although the fitting results from the SIT or SEIT model are similar, the long-term forecast performance can be very different for different models. The results are over 100 experiments.

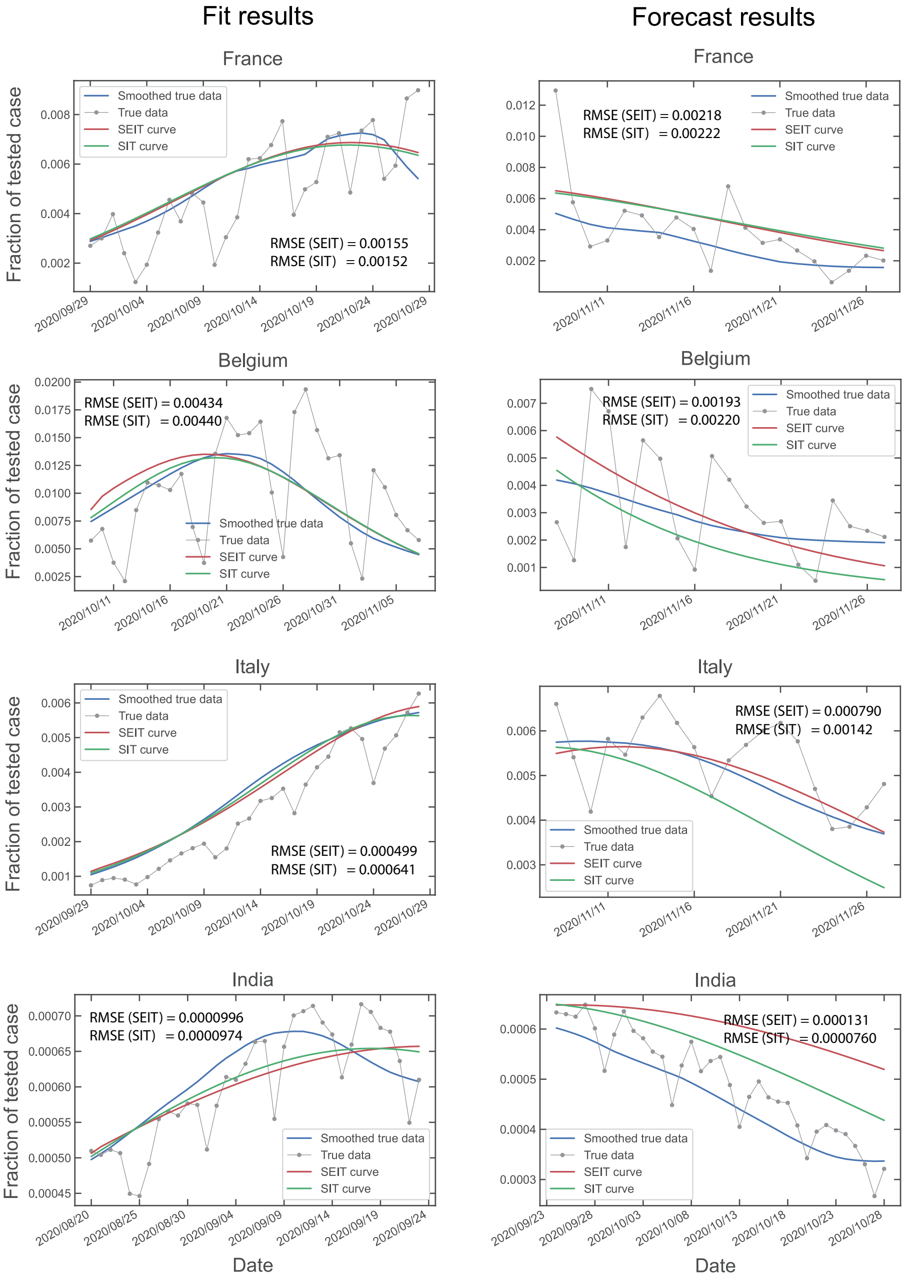

Fig. 6 Fitting and prediction results by SEIT and SIT model on the real data. The real data are from France, Belgium, Italy and India respectively. Before the fitting and forecast, the real data smoothed by moving average with 7-day time window (since there is usually periodicity at around 7 days in real data [25]). It reveals that the fitting results from the SIT and SEIT models are very similar, but the long-term forecast results are significantly different for most countries.

4 Conclusion

Many recent works focus on modeling and forecasting the COVID-19 pandemic based on the compartmental models in epidemiology. A part of these works considered the exposed state, but the others ignored it. This work analyzed to what extend the negligence of the exposed state will affect the fitting and forecast performance. To characterize the COVID-19 pandemic when there are sufficient tests and quarantine, we proposed two revised compartmental models: the Susceptible-Infected-Tested (SIT) model and the Susceptible-Exposed-Infected-Tested (SEIT) model. To characterize the divergence of the forecast performance on SEIT and SIT models, we apply the artificial curves (which are generated by the SEIT model with Gaussian noise) and real data for different countries. We discover that there is no obvious difference between these two models in curve fitting for both the synthetic and real data. However, for the long-term forecast, the prediction results derived from the SIT model and the SEIT model are significantly different for both the synthetic and real data.

References

1. Domenico Cucinotta and Maurizio Vanelli. WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, 91(1):157, 2020.
2. Matheus Henrique Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals*, page 109853, 2020.
3. Bastian Prasse, Massimo A. Achterberg, Long Ma, and Piet Van Mieghem. Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei. *Applied Network Science*, 5(1):1–11, 2020.
4. Massimo A. Achterberg, Bastian Prasse, Long Ma, Stojan Trajanovski, Maksim Kitsak, and Piet Van Mieghem. Comparing the accuracy of several network-based COVID-19 prediction algorithms. *International journal of forecasting*, 2020.
5. Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.
6. Ye Sun, Long Ma, An Zeng, and Wen-Xu Wang. Spreading to localized targets in complex networks. *Scientific reports*, 6(1):1–10, 2016.
7. Davide Faranda and Tommaso Alberti. Modeling the second wave of COVID-19 infections in France and Italy via a stochastic SEIR model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11):111101, 2020.
8. Zeynep Ceylan. Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of The Total Environment*, 729:138817, 2020.
9. Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease*, 12(3):165, 2020.
10. Xiang Zhou, Xudong Ma, Na Hong, Longxiang Su, Yingying Ma, Jie He, Huizhen Jiang, Chun Liu, Guangliang Shan, Weiguo Zhu, et al. Forecasting the worldwide spread of COVID-19 based on logistic model and SEIR model. *medRxiv*, 2020.
11. Ian Cooper, Argha Mondal, and Chris G. Antonopoulos. A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, 2020.
12. Clément Massonnaud, Jonathan Roux, and Pascal Crépey. COVID-19: Forecasting short term hospital needs in France. *medrxiv*, 2020.

13. Armando G.M. Neves and Gustavo Guerrero. Predicting the evolution of the COVID-19 epidemic with the A-SIR model: Lombardy, Italy and Sao Paulo state, Brazil. *Physica D: Nonlinear Phenomena*, 413:132693, 2020.
14. Wangping Jia, Ke Han, Yang Song, Wenzhe Cao, Shengshu Wang, Shanshan Yang, Jianwei Wang, Fuyin Kou, Penggang Tai, Jing Li, et al. Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. *medRxiv*, 2020.
15. Long Ma, Xiao Han, Zhesi Shen, Wen-Xu Wang, and Zengru Di. Efficient reconstruction of heterogeneous networks from time series via compressed sensing. *PloS one*, 10(11):e0142837, 2015.
16. Long Ma, Qiang Liu, and Piet Van Mieghem. Inferring network properties based on the epidemic prevalence. *Applied Network Science*, 4(1):1–13, 2019.
17. Zhihua Liu, Pierre Magal, Ousmane Seydi, and Glenn Webb. A COVID-19 epidemic model with latency period. *Infectious Disease Modelling*, 5:323–337, 2020.
18. Stephen A. Lauer, Kyra H Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.
19. Shiva Moein, Niloofar Nickaeen, Amir Roointan, Niloofar Borhani, Zarifeh Heidary, Shaghayegh Haghjooy Javanmard, Jafar Ghaisari, and Yousof Gheisari. Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan. *Scientific Reports*, 11(1):1–9, 2021.
20. Cory Merow and Mark C. Urban. Seasonality and uncertainty in global COVID-19 growth rates. *Proceedings of the National Academy of Sciences*, 117(44):27456–27464, 2020.
21. Chris Dye and Nigel Gay. Modeling the SARS epidemic. *Science*, 300(5627):1884–1885, 2003.
22. Haiping Fang, Jixiu Chen, and Jun Hu. Modelling the SARS epidemic by a lattice-based Monte-Carlo simulation. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7470–7473. IEEE, 2006.
23. Can Hou, Jiaxin Chen, Yaqing Zhou, Lei Hua, Jinxia Yuan, Shu He, Yi Guo, Sheng Zhang, Qiaowei Jia, Chenhui Zhao, et al. The effectiveness of quarantine of wuhan city against the corona virus disease 2019 (COVID-19): A well-mixed SEIR model analysis. *Journal of medical virology*, 2020.
24. Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error over the root mean square error in assessing average model performance. *Climate research*, 30(1):79–82, 2005.
25. Aviv Bergman, Yehonatan Sella, Peter Agre, and Arturo Casadevall. Oscillations in US COVID-19 incidence and mortality data reflect diagnostic and reporting factors. *Msystems*, 5(4), 2020.

A Algorithm

A.1 Hill climbing optimization

Hill climbing is one of the heuristic mathematical optimization method, attempting to find the maximizer or the minimizer of a function. Here in our case, we implement this algorithm and try to find the optimal parameters to fit some curves and produce root-mean-square-error (RMSE) as small as possible.

The basic idea of hill climbing is that the algorithm accepts any modifications regarding the variables that can optimize the cost function value. To be specific, if one would like to minimize the targeted cost function, one can search all the neighbours of the current state or randomly generate a new state, and if the corresponding cost function value decreases, with the hope that this direction will lead to the global optima, the new state is accepted. Although this algorithm can be illustrated that the eventual solution is not necessarily the global optima, yet due to the complicated relationship between parameters τ, β and γ and the corresponding generated curve, this method is suitable for this fitting case. Furthermore, as shown in the following section, the results are quite good, indicating the efficacy of the hill climbing algorithm. The implementation of this algorithm can be followed by the pseudo code in the next section.

A.2 Pseudo code

In this section, we are going to show the way to implement hill climbing algorithm in Algorithm 2.

Algorithm 1: Hill Climbing Algorithm for the artificial data

```
Data: $T_{noise}$ data created by SEIT model with additive Gaussian Noise
Result: Optimal $\tau$, $\beta$, and $\gamma$ with the lowest $RMSE$
step size $\leftarrow$ 1, iteration $\leftarrow$ 5000
$\beta_0 \leftarrow$ randomly drawn from uniform distribution (0, 1), $\tau_0 \leftarrow$ randomly drawn
  from uniform distribution (0, 5), $\gamma_0 = \beta_0/\tau_0$
Let state $s = s_0$, the initial state of an SEIT or SIT curve which is created by $\beta_0$
  and $\gamma_0$
Energy $E(s)$, defined as RMSE between $s$ and $T_{noise}$
for $n$ in iteration do
    $s_{new} \leftarrow neighbor(s)$ generated by adding uniformly distributed variable in
      (-1,+1) × step size
    if $E(s) > E(s_{new})$ then
        $s \leftarrow s_{new}$
    end
end
Return $\tau$, $\beta$ ,$\gamma$
```

Algorithm 2: Hill Climbing Algorithm for the real data

```
Data: $T_{daily}$ Real daily-increased data from various countries after smoothing
Result: Optimal $\tau$, $\beta$, and $\gamma$ with the lowest $RMSE$
step size $a_1 \leftarrow$ 1, step size $a_2 \leftarrow$ 0.0001, iteration $\leftarrow$ 50000
$\beta \leftarrow$ randomly drawn from uniform distribution (0, 1), $\tau \leftarrow$ randomly drawn from
  uniform distribution (0, 5), $\gamma = \beta/\tau$
$y_0 \leftarrow [S_0,(E_0,)I_0,T_0]$
Let state $s = s_0$, the initial state of an SEIT or SIT curve which is created by $\beta$, $\gamma$
  and $y_0$
Energy $E(s)$, defined as RMSE between $s$ and $T_{daily}$
for $n$ in iteration do
    $s_{new} \leftarrow neighbor(s)$ generated by adding uniformly distributed variable in
      (-1,+1) × $a_1$ to $\beta$, $\tau$ and (-1,+1) × $a_2$ to $T_0$
    if $E(s) > E(s_{new})$ then
        $s \leftarrow s_{new}$
    end
end
Return $\tau$, $\beta$ ,$\gamma$
```