COVID-19 dynamics across the US: A deep learning study of human mobility and social behavior

Mohamed Aziz Bhouri, Francisco Sahli Costabal, Hanwen Wang, Kevin Linka, Mathias Peirlinck, Ellen Kuhl, Paris Perdikaris

Abstract

This paper presents a deep learning framework for epidemiology system identification from noisy and sparse observations with quantified uncertainty. The proposed approach employs an ensemble of deep neural networks to infer the time-dependent reproduction number of an infectious disease by formulating a tensor-based multi-step loss function that allows us to efficiently calibrate the model on multiple observed trajectories. The method is applied to a mobility and social behavior-based SEIR model of COVID-19 spread. The model is trained on Google and Unacast mobility data spanning a period of 66 days, and is able to yield accurate future forecasts of COVID-19 spread in 203 US counties within a time-window of 15 days. Interestingly, a sensitivity analysis that assesses the importance of different mobility and social behavior parameters reveals that attendance of close places, including workplaces, residential, and retail and recreational locations, has the largest impact on the effective reproduction number. The model enables us to rapidly probe and quantify the effects of government interventions, such as lock-down and re-opening strategies. Taken together, the proposed framework provides a robust workflow for data-driven epidemiology model discovery under uncertainty and produces probabilistic forecasts for the evolution of a pandemic that can judiciously provide information for policy and decision making. All codes and data accompanying this manuscript are available at https://github.com/PredictiveIntelligenceLab/DeepCOVID19.

Keywords: Epidemiology model discovery; Sensitivity analysis; Neural networks; Scientific machine learning; Dynamical systems; Uncertainty quantification

© 2021 Elsevier B.V. All rights reserved.

https://doi.org/10.1016/j.cma.2021.113891
0045-7825/© 2021 Elsevier B.V. All rights reserved.
1. Introduction

1.1. Background

Since its emergence in late December 2020 [1], as of early September 2020, the novel SARS-CoV-2 virus (COVID-19) has resulted in over 81 million confirmed cases globally, of which nearly 2 million people have died [2]. After nearly months, active cases still surge in many countries and regions [2]. To date, there are no known preventive pharmaceutical interventions [3], and numerous behavioral intervention policies, such as lockdown, mandatory self-isolation and facial covering orders [4] have been issued to reduce the transmission of such a highly infectious disease to an extent that current health and financial systems could sustain. Without these timely countermeasures, breakdown of these systems will lead to more severe social and economic crisis in an unexpectedly fast pace, especially given the exponential rise in cases number at the beginning of the outbreak as it was the case in the US, for instance for the first few months of the COVID-19 spread [5]. Despite their effectiveness in containing the aggressive virus spread, these attempts have also interrupted ordinary living orders, and have resulted in heavy socio-economical impacts such as exploding unemployment rates, prevailing emotional distress, and the consequently exacerbated civil unrest [6] that has evidently encouraged super spreader events. Phased re-opening strategy [7], as adopted by many states and cities as a compromise, seemingly becomes the only way to leverage the casualties brought by the disease itself, and by the consequential economic interruption. Therefore, accurately and quantitatively assessing the cost and benefit of restricted opening gains an increasing importance nowadays, especially at a time when America saw spikes after spikes in tested positive population after some controversial policies.

Studies [8,9] have revealed that the transmission of COVID-19 likely comes from direct contact, such as droplets from infected persons, or indirect contact, where virus that remained active on shared surface is brought into human body by behaviors such as whipping face with hands [10]. The virus replicates itself in the throat, and further causes lung lesions. The incubation period, before human immune responses begin to intervene, could last 6 days on average or even as long as 14 days observed [11]. During or even after the incubation period, the patients can still remain largely asymptomatic, while being able to infect others [12–14]. The corresponding antigen detection assays using throat wash and saliva are also reported to produce false result [15], affecting the reliability of the case data resource.

Such pathology makes clear that close contact interactions in public space is a major cause of the spread of COVID-19. Individual level transmission models [16] have shown that 80% of secondary transmissions may have been caused by a small fraction of infectious individuals (∼10%) from some super-spreading events, mainly large and dense gatherings. The aforementioned two challenges, long incubation period and noisy data, call for timely adjustments to existing policy plans based on robust assessments of the situation. Under such motivation, technology companies such as Google [17] and Apple [18] also make publicly available their privacy preserved databases.

1.2. Related work

Multi-compartment models, such as SEIR and SIR [19], have been used extensively in the infectious disease analysis and prediction. Over the years, multi-compartment models have been also been augmented to account for more specific population groups that are subject to different transmission rates, recovery rates, etc. [20]. However, little attention has been received towards extending multi-compartment models to account for public mobility and behavioral data that can reflect non-pharmaceutical interventions, such as government policies, daily migration and interaction patterns, that could also contain a rich information about the disease transmission. Recent studies have attempted to account for spatial correlations to model the transfer of infected population between different geographic regions by utilizing human travel data from transportation and customs authorities [5,21,22]. Despite their novelty, such models also bring in questionable assumptions. For example, some highly culturally directed migration may not resemble a diffusion process described by a partial differential equation model, and the unrecorded traveling by foot or by personal vehicles could bring unique biases into the error of the travel data. In conjunction with the aforementioned works, time- and mobility-dependent parameters are also being introduced in compartment models to enable the incorporation of more diverse data sources [20].

As multi-compartment models become more sophisticated to reflect the complexity of the COVID-19 pandemic, a need for fast and reliable parameter estimation methods arises. Classical approaches to data assimilation, including
Bayesian methods [23–25] and Kalman filters [26], offer computational infrastructure for fitting multi-compartment models to observational data. However, their use can be often hindered by prerequisites for model construction, such as linearity [27], and rapidly increasing computational complexity as the number of model parameters to be inferred and the size and diversity of the training data-sets increase. Therefore, there is a need for robust and scalable computational frameworks for epidemiology model calibration from noisy and incomplete time-series data.

1.3. Contributions of this work

Here we conduct a large-scale computational study that leverages Google’s mobility trend data [17] around grocery and pharmacy stores, parks, public transportation, retail and recreation spaces, residential and workplaces, along with the Unacast’s social behavior data [28] which includes information on the average distance traveled, visits to non-essential retail and services, and unique human encounters per km$^2$ relative to national metrics pre-COVID-19. We believe that these types of mobility and social behavior trends cover most scenarios of social mixing and can be used as predictive indicators of viral spread [29].

From a modeling perspective, the main novelty of this work stems from utilizing mobility and social behavior trends data to model the temporal evolution of the COVID-19 infectious rate — the probability of transmitting disease between a susceptible and an infectious individual (Fig. 1). From a methodological standpoint, this work generalizes the multi-step neural network method of Raissi et al. [30] to enable gradient-based optimization of unknown epidemiology model parameters with latent variables. This is achieved by effectively back-propagating the influence of the unobserved variables through the disease spread dynamics with an appropriate tensorization of the loss function. This key step enables a computationally efficient treatment of multiple temporal trajectories, as well as the concurrent training of multiple neural networks. The latter, allows us to generate a statistical ensemble over the predicted evolution of the underlying disease spread dynamics, which enables future forecasting with quantified uncertainty.

In summary, the specific contributions of this work can be organized as follows:

- We extend the multi-step neural network method of Raissi et al. [30] to enable dynamical system’s parameters inference with unobserved latent variables. Moreover, we tensorize the loss function formulation in order to perform the parameters inference from multiple trajectories of different counties, while having a small memory footprint in terms of the size of the associated computational graph. Hence, the proposed approach is computationally efficient and enables training of multiple neural networks, which provides an uncertainty quantification estimate of the dynamical system evolution via multiple trajectory forecasts.
- We model the effective reproduction number with a long short-term memory neural network (LSTM) to capture its temporal dependence on the mobility and social behavior variables. The LSTM architecture enables to account for the time lag between the change in mobility and social behavior, and the effect on the spread of the disease.
- We demonstrate the effectiveness of the proposed method by applying it to the COVID-19 spread in 203 US counties such that the training is performed using data for a period of 66 days, and we produce reliable future forecasts with quantified uncertainty for a period of 15 days. The lock-down efficiency is quantified by the subsequent forecasted decrease of the effective reproduction number. We also perform a sensitivity analysis showing that among the mobility parameters, the attendance of closed places has the highest influence on the evolution of the effective reproduction number, unlike open or transit locations. Moreover, all social behavior parameters have relatively important sensitivities, highlighting the importance of social interactions on the spread of the virus.

Our findings put forth a novel, flexible and robust workflow for data-driven epidemiology model discovery that can potentially help health authorities take a more quantitative and judicious approach to policy making in order to slow down COVID-19 propagation. We demonstrate that well-calibrated epidemiology models can indeed return sensible future forecasts. Such capability can directly assist with optimizing resource allocation for more precise diagnostics and treatment, as well as yield a better understanding of the disease spread dynamics. Taken together, this can be another tool in our arsenal to help us better prepare for COVID-19 and potential future pandemics, and provide additional quantitative tools to help us minimize their spread.

This paper is organized as follows. Section 2.1 presents the mobility and social behavior-based SEIR epidemiology model considered to study the spread of COVID-19. Section 2.2 details the proposed computational method.
Our goal is to develop a robust computational framework for calibrating COVID-19 models from noisy time-series data. This will enable the quantitative assessment of government policies (e.g., travel bans, quarantine, contact tracing, etc.), and yield reliable forecasts with quantified uncertainty that can help us manage and contain an epidemic.

Inference of the mobility and social behavior-based SEIR model parameters and the corresponding technical ingredients. In Section 3, the effectiveness of the proposed method is demonstrated by testing it on real mobility and infections data from 203 US counties and obtaining reliable future forecasts for a period of 15 days with quantified uncertainty. Within the same section, we present a sensitivity analysis to assess the influence of the different mobility and social behavior parameters considered on the spread of the disease. Finally, in Section 4 we summarize our key findings, discuss the limitations of the proposed approach, and carve out directions for future investigation.

2. Methods

2.1. Epidemiology model

We consider a mobility and social behavior-based SEIR model detailed as follows:

\[
\begin{align*}
\frac{dC}{dt} &= \frac{dI}{dt} + \gamma I \\
\frac{dS}{dt} &= -\beta(m(t)) S I \\
\frac{dE}{dt} &= \beta(m(t)) S I - \delta E \\
\frac{dI}{dt} &= \delta E - \gamma I \\
\frac{dR}{dt} &= \gamma I,
\end{align*}
\]

where

- \(S\) is the fraction of susceptible population,
- \(E\) the fraction of exposed population including individuals who have been infected but are not yet infectious themselves during the incubation period,
- \(I\) the fraction of infectious population,
- \(R\) the fraction of removed population that accounts for recovered and deceased individuals, and
- \(C(t) = R(t) + I(t)\) is the fraction of cumulative cases, which is the only variable with reliable data over time.

The only available data-sets at the early stages of the COVID-19 pandemic consist of the cumulative cases \(C(t)\), the deaths and the recoveries. In the SEIR model, the removed population \(R\) accounts for the deaths and the recoveries. However, the recoveries are never accurately measured since, for the positively diagnosed patients, only those in serious or critical condition are closely followed by healthcare providers. Hence, we can accurately determine when such patients move from the \(I(t)\) to the \(R(t)\) compartment. However, such patients in serious or critical condition represent a very small fraction of the infected population \(I(t)\) (for instance, they only represent around 0.4% of the positively diagnosed patients these days [https://www.worldometers.info/coronavirus/]), while most of these patients are asked to stay at home and then are tested after two weeks or so in order to verify if they are still infectious or not. As a consequence, the exact day of recovery for most of the infected people is not accurately known, which leaves us only the cumulative cases as the most reliable indicator. The same reasoning applies with the disparity in reporting the infected population in urban and rural areas, either due to undetected or asymptomatic patients. Such population can have an influence on the model behavior when the case numbers are relatively large. However, estimations of the undetected population based on antibody studies are currently available only at very few locations and only...
in urban areas [31]. For this reason, we decided to not explicitly model this population, as we would be simply guessing its size. Note that we consider an SEIR model with variables normalized by the number of population such that the resulting model (1) is universal for all regions and independent of the number of their inhabitants.

Since the spread of an infectious disease that is transmitted from human to human depends on interpersonal interactions, we model the susceptible to infectious transition rate $\beta$ in Eq. (1) as a function of mobility and social behavior parameters denoted as $m \in \mathbb{R}^{N_m}$, which contains the mobility and social behavior parameters $m_1(t), \ldots, m_{N_m}(t)$. This enables determining the evolution of the effective reproduction number, $R_0(t) \equiv \frac{\beta(m(t))}{\gamma}$, a key parameter to characterize the evolution of an epidemic.

The proposed model is trained on data provided by Google for mobility change from baseline [17] (parameters $m_1, \ldots, m_6$ as discussed below), along with the data provided by Unacast [28] for social behavior trends (parameters $m_7, \ldots, m_9$ as detailed below) for all 203 US counties considered in this work. Specifically, we consider a total of $N_m = 9$ mobility and social behavior parameters corresponding to:

- $m_1$: retail and recreation percent change from baseline,
- $m_2$: grocery and pharmacy percent change from baseline,
- $m_3$: parks percent change from baseline,
- $m_4$: transit stations percent change from baseline,
- $m_5$: workplaces percent change from baseline,
- $m_6$: residential percent change from baseline,
- $m_7$: average distance traveled,
- $m_8$: visits to non-essential retail and services,
- $m_9$: unique human encounters per Km$^2$ relative to national pre-COVID-19.

Fig. 2 provides the observed mobility and social behavior trends for three counties randomly selected among the data-set considered, showing the inherent presence of noise within the trajectories.

2.2. A deep learning approach to model calibration given partially observed states

In this work, $\gamma$ and $\delta$ are assumed as given and equal to 1/6.5 [32,33] and 1/2.5 [11,32] respectively for COVID-19. Additional numerical studies that aim to elucidate the model’s sensitivity on $\gamma$ and $\delta$ have been conducted. Our findings indicate that the specific values chosen for $\gamma$ and $\delta$ do not have a significant impact on the model’s predictions as shown in Appendix C). One of our main goals is to learn the dynamics of $\beta(t)$, and,
consequently, the evolution of the solution to the SEIR model of Eq. (1). The learning task can then be formulated as follows.

Given $\gamma$, $\delta$ and a data-set $\mathcal{D}$ for $N_c$ different regions containing: (i) the initial conditions of the variable $I(t = t_0)$ for each region, (ii) the time trajectories of the observed variable $C(t)$ for each region, and (iii) the time trajectories of the observed mobility and social behavior trends $m(t)$ for each region, infer the dependence of $\beta(t) \equiv \beta(m(t))$ and the initial fractions of the exposed population $E(t = t_0)$ for all the different regions.

The observed data-set $\mathcal{D}$ can be expressed as follows:

$$\mathcal{D} \equiv \{ I^1(t_0), \ldots, I^{N_c}(t_0), C^1(t_0), \ldots, C^{N_c}(t_0), \ldots, C^1(t_{N_t-1}), \ldots, C^{N_c}(t_{N_t-1}), m^1(t_{t-1}), \ldots, m^{N_c}(t_{N_t-2}), \ldots, m^1(t_{t-2}), \ldots, m^{N_c}(t_{N_t-1}) \}, \quad (2)$$

where superscripts refer to the different regions considered. $N_t$ refers to the number of days for which the number of cumulative cases is available. Those time instances are represented with $t_0, t_1, \ldots, t_{N_t-1}$. In order to learn the mapping $\beta \equiv \beta(m(t))$, $\beta$ is modeled with a long-short term memory (LSTM) [34] neural network of the form

$$\hat{\beta} = f(m, \theta_{NN}), \quad (3)$$

where $\theta_{NN}$ refers to the weights and bias parameters defining the LSTM network. Since we are modeling the spread of an infectious disease with an incubation period, we utilize the lag structure of the LSTM network to model $\beta$ as a function of the mobility and social behavior parameters over a period of time of $\tau$ days. Hence, we model the susceptible to infectious transition rate $\beta(t_i)$ as a function of the parameters $m$ on the $\tau$ previous days: $m(t_{t-\tau}), m(t_{t-\tau+1}), \ldots, m(t_{t-1})$. As a consequence, to model the cumulative cases from $t_0$ to $t_{N_t-1}$, and hence the susceptible to infectious transition rate $\beta$ from $t_0$ to $t_{N_t-1}$, we need to consider the parameters $m$ from $t_{t-\tau}$ to $t_{N_t-2}$ as given in Eq. (2). Note that due to the incubation period and the testing delay, we considered that the parameters $m$ at a given day $t_i$ could not influence the propagation of the disease that is measured and reported on the exact same day $t_i$, which justifies the one day shift in the ending day of the ranges considered for the cumulative cases and the parameters in the data-set $\mathcal{D}$ needed to train the model. The LSTM takes care of capturing the temporal dependence of $\beta$ on the mobility and social behavior parameters $m$ within the $\tau$ previous days. The forward pass of LSTM neural network considered in this work is detailed in Appendix A.

Given the available dataset $\mathcal{D}$, we formulate a multi-step loss function in order to infer the learnable parameters $\theta$ given in Eq. (4), which include both the LSTM parameters $\theta_{NN}$ as well as the latent initial fractions of the exposed population for all $N_c$ different regions considered,

$$\theta \equiv (\theta_{NN}, E^1(t_0), \ldots, E^{N_c}(t_0)). \quad (4)$$

To formulate the multi-step loss function, we follow the approach put forth by Raissi et al. [30], and consider an appropriate temporal discretization of Eq. (1). Given the available data containing the initial conditions of the variable $I(t = t_0)$ and the trajectories of the variable $C(t)$, the variable $I(t)$ can be estimated for the different time instances $t_1, \ldots, t_{N_t-1}$ using the first equation of the SEIR model (1). For instance, using the trapezoidal discretization rule [30,35], the first equation of the SEIR model (1) for region $j$ gives:

$$I^j(t_0) = C^j(t_0), \quad [F + T] I^j = F C^j + \begin{bmatrix} -\gamma C^j(t_0) \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (5)$$

where

$$F = \frac{1}{dt} \begin{bmatrix} 1 & -1 & \cdots & 0 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{N_t-1 \times N_t-1}, \quad T = \begin{bmatrix} 1/2 & 1/2 & \cdots & 1/2 \\ 1/2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1/2 \\ 1/2 & \cdots & 1/2 & 1/2 \end{bmatrix} \in \mathbb{R}^{N_t-1 \times N_t-1}, \quad (6)$$

dt is the time-step size corresponding to the period of the infections number update, and taken equal to 1 day in this work based on the available data, and $C^j, I^j \in \mathbb{R}^{N_t-1}$ are the vectors containing $C^j(t)$ and $I^j(t)$ at the time instances $t_1, \ldots, t_{N_t-1}$, respectively.
Then, given the estimated \( I(t) \) from Eq. (5) and the parameters \( \theta \), we can compute \( \hat{S}(t) \), then \( \hat{E}(t) \) and finally \( \hat{I}(t) \) using the remaining equations of the SEIR model (1). Using the trapezoidal rule for instance, such discretization for region \( j \) gives:

\[
[F + \hat{B}] \hat{S}^j =
\begin{bmatrix}
\hat{S}^j(t_0) = 1 - E^j(t_0) - I^j(t_0),
\frac{1 - E^j(t_0) - C^j(t_0)}{dt} - \frac{\hat{\beta}_j(m^j, \theta_{NN}) (1 - E^j(t_0) - C^j(t_0)) C^j(t_0)}{2}
\end{bmatrix},
\]

\[
[F + \delta \, T] \hat{E}^j = \hat{B} \hat{S}^j +
\begin{bmatrix}
E^j(t_0) + \frac{\hat{\beta}_j(m^j, \theta_{NN}) (1 - E^j(t_0) - C^j(t_0)) C^j(t_0)}{2} - \delta E^j(t_0)
\end{bmatrix},
\]

\[
[F + \gamma \, T] \hat{I}^j = \delta \, T \hat{E}^j +
\begin{bmatrix}
\frac{C^j(t_0)}{dt} + \frac{E^j(t_0)}{2} - \gamma \frac{C^j(t_0)}{2}
\end{bmatrix},
\]

where

\[
\hat{B} =
\begin{bmatrix}
\frac{\hat{\beta}_1 I^j(t_1)}{2} & \cdots & \frac{\hat{\beta}_1 I^j(t_1)}{2} \\
\vdots & \ddots & \vdots \\
\frac{\hat{\beta}_{N_t-2} I^j(t_{N_t-2})}{2} & \cdots & \frac{\hat{\beta}_{N_t-2} I^j(t_{N_t-2})}{2}
\end{bmatrix}\in\mathbb{R}^{N_t-1 \times N_t-1},
\]

\( \hat{\beta}_j \equiv \hat{\beta}(m^j(t_{i-1}), \ldots, m^j(t_{i-1}), \theta_{NN}) \) is the LSTM prediction of \( \beta \) for region \( j \) at time instance \( t_i \), and \( \hat{S}^j, \hat{E}^j, \hat{I}^j \in \mathbb{R}^{N_t-1} \) are the vectors containing \( \hat{S}(t_i), \hat{E}(t_i) \) and \( \hat{I}(t_i) \) at the time instances \( t_1, \ldots, t_{N_t-1} \), respectively. Note that \( E^j(t_0) \) appearing in the initial conditions in Eqs. (7) is part of the learnable parameters \( \theta \), as described in Eq. (4).

Notice that, given the dataset \( D \) and the parameters \( \theta \), one can estimate \( \hat{S}(t), \hat{E}(t), \) and \( \hat{I}(t) \) in that order. Finally, the loss function for a single neural network is expressed as the difference between the variable \( I(t) \) estimated only from the available data and the estimated variable \( \hat{I}(t) \) which depends on the learnable parameters:

\[
\mathcal{L}(\theta) = \frac{1}{(N_t - 1) \times N_c} \sum_{i=1}^{N_t-1} \sum_{j=1}^{N_c} [I^j(t_i; D) - \hat{I}^j(t_i; \theta)]^2.
\]

One crucial point in terms of implementation consists of the possibility of expressing the finite difference equation of (5), and each of the finite difference equations of (7), as tensors of shape \( N_c \times (N_t - 1) \). As a consequence, the loss \( \mathcal{L} \) given in Eq. (9) can be expressed as the sum of the elements of a tensor of shape \( N_c \times (N_t - 1) \) ensuring a minimalist computational graph, which speeds up gradient back-propagation. We emphasize that the proposed methodology is independent of the chosen finite difference scheme, and the trapezoidal rule is chosen here for sake of presentation and in coherence with the finite difference scheme adopted in this work as justified in Appendix A. This computational strategy allows us to consider multiple trajectories simultaneously (accounting for the different regions considered in the data), instead of performing the learning task using observations from a single or a handful of trajectories. Finally, the loss can be further tensorized to enable the parallel training of multiple neural networks.
By training an ensemble of independent networks, we can obtain a statistical characterization of model uncertainty and quantify its effect on the resulting model predictions and forecasts. For more technical details, the interested reader is referred to our publicly available implementation.

3. Results

All code and data presented in this section will be made publicly available at https://github.com/PredictiveIntelligenceLab/DeepCOVID19. The minimization of Eq. (9) is carried out using stochastic Adam updates [36] with the hyper-parameters summarized in Table 1. Based on our experience, the reported results are robust with respect to the numerical values chosen for the different optimization hyper-parameters.

### 3.1. Training data and pre-processing

We remove the counties containing missing mobility or social behavior parameters for more than 3 consecutive days or 20 days in total over the considered time period, and we use linear interpolation for missing data for 3 or less consecutive days. We also consider the USAFacts processed infections data for all US counties [37]. The two datasets are cross-matched together such that the considered counties satisfy the following criteria:

- the infection rate $C(t)$ is at least equal to 0.2% after $t = 81$ days of the spread of the disease,
- the mobility and social behavior data for the $\tau$ days prior to the first day of infections is available.

For each county, the first day of infections is defined as the first day for which there are at least 10 infections. These criteria provide a dataset with 203 US counties and a period of time of 81 days, such that the first day of infections for each county is between March 16, 2020 and March 24, 2020, while the last day of considered is between June 4, 2020 and June 12, 2020. The total population of these counties is 150 million, which represents 46% of the total US population.

### 3.2. Predicted forecasts

The proposed framework is trained on the collected data corresponding to the first 66 days. For the remaining 15 days, only the mobility and social behavior data are used in order to extrapolate the fraction of infected population. We train an ensemble of 100 models with randomized parameter initializations. The total training time takes 14.7 min on a single NVIDIA Tesla P100 GPU. Once trained, the model provides estimates of the cumulative cases fraction with a relative error equal to 4.3% on the training data (for the first 66 days), and a relative error equal to 6.6% on the testing data, corresponding to the extrapolation carried out for the remaining 15 days of the data-set. Fig. 3 shows the distribution of relative errors averaged over the 15 days extrapolation window across all US counties considered. Evidently, the proposed model can yield accurate future forecasts of COVID-19 spread, particularly for counties having a higher fraction of cumulative cases. However, using an SEIR model with a constant $\beta$ over time for each county fails to explain the spread of COVID-19 for the 203 US counties considered. Fig. B.13 in Appendix B compares the relative forecasting errors averaged over the extrapolation period when using the proposed multi-step LSTM approach and a multi-step approach with a constant susceptible to infectious transition rate $\beta$. Based on these results we conclude that enhancing the expressiveness of the SEIR model via the inclusion

---

**Table 1**

Hyper-parameter settings employed across all numerical studies.

<table>
<thead>
<tr>
<th>Hyper-parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{NN}$ initialization</td>
<td>Xavier initialization</td>
</tr>
<tr>
<td>$E(t_0), \ldots, E(t_N)$ initialization</td>
<td>Uniform initialization</td>
</tr>
<tr>
<td>Optimization algorithm</td>
<td>Adam</td>
</tr>
<tr>
<td>Learning rate</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>$2 \times 10^4$</td>
</tr>
<tr>
<td>County batch size</td>
<td>60</td>
</tr>
</tbody>
</table>
of mobility and social behavior-based effects in the time-depended modeling of $\beta$ has a significant impact on the generalization accuracy of the model, leading to more accurate future forecasts.

Fig. 4 provides the dependency of the relative error averaged over the 15 days extrapolation on the fraction of cumulative cases on the last day. The latter shows a better accuracy for counties with higher cumulative cases fractions, which can be explained by the loss function considered (9). Indeed, counties with higher fraction of cumulative cases present a higher fraction of the infectious population. As a consequence, the corresponding terms within the loss function are generally higher than the ones related to counties with a lower fraction of cumulative cases. Hence the learning of the model is inherently more driven towards minimizing the errors of the estimates related to counties with higher fraction of cumulative cases, which explains the lower errors observed compared to counties with lower fraction of cumulative cases. One possibility to remedy to this issue is to consider normalized errors in the loss function (9) by dividing the L2 errors $\left[I^j(t_i) - \hat{I}^j(t_i)\right]^2$ by the values of $I^j(t_i)$, $1 \leq i \leq N_t - 1$, $1 \leq j \leq N_c$ added with a zero or positive hyper-parameter $\epsilon_L$ that needs to be tuned in order to prevent significantly small values in the denominator. However, such approach drove the model to better approximate counties with lower fraction of cumulative cases due to the importance weight introduced by the corresponding small denominator $I^j(t_i)$. Since accurately modeling counties with higher disease propagation, and hence higher fraction of cumulative cases, is more crucial than focusing on counties with lower infection rates, we opted for the non-normalized loss function. Note that considering the normalized loss partially remedied the disparity in accuracy for the different counties, however the improvement observed was not significant regarding the prediction and extrapolation errors and it required a fine tuning of the hyper-parameter $\epsilon_L$, which would make the approach a problem-dependent method since $\epsilon_L$ would differ depending on the data-set considered.

New York state has the highest number of counties with a fraction of infected population above 1% after 81 days of the spread of the virus. Fig. 5 shows the predictions obtained for the evolution of the fraction of infected population for the major New York counties. Figs. 6 and 7 provide additional sample results for other US counties. New Jersey and Massachusetts present the second and third highest number of counties with a fraction of infected population above 1% after 81 days of the spread of the virus.

### 3.3. Sensitivity analysis

In order to assess the influence of different mobility and social behavior parameters on the evolution of the susceptible to infectious transition rate $\beta$, a sensitivity analysis using the Morris method [38] is performed on the

---

1 Predictions for a longer window of 21 days are also provided in Appendix D.
trained LSTM neural network predicting $\hat{\beta} = f(\mathbf{m}, \theta_{NN})$. Note that the LSTM has $N_m \times \tau$ inputs, such that $\hat{\beta}(t_k)$ inference is carried out using $\mathbf{m}(t_k - \tau), \ldots, \mathbf{m}(t_k - 1)$, where we remind the reader that $\mathbf{m}(t) \in \mathbb{R}^{N_m}$ contains the mobility and social behavior parameters $m_1(t), \ldots, m_{N_m}(t)$. We use $s_{i,j}$ to denote the modified mean of the Morris method sensitivity measure \cite{39} of $\hat{\beta}(t_k)$ to $m_i(t_k - j)$.

Fig. 8 presents a heat-map of the normalized sensitivity measures $s_{i,j}$. As expected, in view of the Google mobility parameters ($m_1, \ldots, m_6$), these results suggest that the sensitivity of $\beta$ to parks attendance (parameter $m_3$) — the only open place that is considered among the six available mobility parameters — is the lowest amongst all six. The second least influential parameter is the transit station’s attendance (parameter $m_4$), reflective of a location where people spend on average the least amount of time among the places considered in the remaining five mobility parameters (except parameter $m_3$). The numerical values obtained are also consistent with the COVID-19 incubation period estimated to be between 2 and 21 days by the United States’ CDC \cite{40}. Indeed, the mobility parameters do not instantaneously affect $\beta$ and a lag period should be taken into account in order to observe the effects of the mobility parameters on the spread of the disease, which corresponds to the incubation period and the testing delay. Regarding the Unacast social behavior parameters, all parameters display a significant sensitivity, with unnecessary visits displaying the highest value as shown in Fig. 8.

We also perform a local sensitivity analysis as follows. We increase each of the mobility and social behavior parameters by 10% for the last 21 days (each at a time), and we determine the corresponding prediction for the susceptible to infectious transition rate $\beta$. We then compare the obtained value for $\beta$ for the last day to the actual one obtained with the unchanged data. Finally, we pick the parameter that corresponds to the highest change in $\beta$. This result, shown in Fig. 9, allows us to locally identify the most sensitive parameter among the mobility and social behavior ones considered for each of the 203 counties studied. Interestingly, our results suggest that the spread of COVID-19 among a population of over 60 million people is most sensitive to the closure and reopening of workplaces, followed by changes in visits to residential places.

3.4. Effects of lock-down

In Figs. 5–7 we indicated the day on which the $\geq 50$ gatherings ban was issued along with the stay at home order day (except for Navajo, Arizona for which the $\geq 50$ gatherings ban was issued at day $t = -4$). The effect of these interventions on the evolution of the effective reproduction number $R_0(t)$ is significant, as the latter generally drops following at least one of these two policy decisions, such that $R_0(t)$ always reaches values around 1 while these decisions are still in place. We should note that banning $\geq 50$ gatherings was sufficiently effective to decrease $R_0$ in some locations, including most of New York counties; Massachusetts counties; and Prince George’s, Maryland, for instance. However, for other counties, the drop of $R_0$ was significant only after issuing the stay at home order as observed for Suffolk, New York; Middlesex, New Jersey; Cook, Illinois; Jefferson, Louisiana; and Prince Williams, Virginia. These observations quantify the importance of limiting gatherings in managing COVID-19 spread and the
Fig. 5. Evolution of the fraction of cumulative cases $C(t)$ and the effective reproduction number $R_0(t)$ across different New York state counties. The orange and green vertical dashed lines indicate the dates corresponding to the large crowd gathering ban and stay at home order issued by each county. The pink vertical dashed line delineates the data used for training the model (days 0–65), and the predicted model forecasts (days 66–80).
Fig. 6. Evolution of the fraction of cumulative cases $C(t)$ and the effective reproduction number $R_0(t)$ for representative counties across the US. The orange and green vertical dashed lines indicate the dates corresponding to the large crowd gathering ban and stay at home order issued by each county. The pink vertical dashed line delineates the data used for training the model (days 0–65), and the predicted model forecasts (days 66–80).
necessity of issuing the stay at home orders in locations where gatherings bans alone are not enough to manage the spreading.

3.5. Effects of re-opening

To quantify the effect of re-opening, the retail and recreation mobility parameter of New York county (New York) is artificially altered as shown in top left plot of Fig. 10 (while the other mobility and social behavior parameters...
Fig. 8. $s_{i,j}$: Sensitivity of susceptible to infectious transition rate $\beta$ to the mobility and social behavior parameters.

Fig. 9. Mobility and social behavior parameters with the greatest effect on the reproduction number, by county (left panel) and by population (right panel).

are kept the same. As shown in top right and bottom left plots of Fig. 10, such policy has a significant effect on the evolution of $\beta(t)$ and by consequence on the effective reproduction number $R_0(t)$ increased beyond 1 for the last 20 days of the simulation. As a result, the number of cumulative cases $C(t)$ would significantly increase as shown in the bottom right plot of Fig. 10 and the spread of the disease would be accelerated due to the increase of the retail and recreation mobility parameter. We note that the reproduction number is significantly affected only after 20 days of increasing the mobility parameter. This result is in line with the sensitivity analysis in Fig. 8, which shows that the reproduction number is mostly affected by mobility between 17 and 21 days earlier. In practice, this delay poses a challenge to policy makers, which can only evaluate the effects of easing and imposing restrictions after two to three weeks.

Moreover, the altered trajectories show a higher uncertainty than the baselines in Fig. 10. The predictive uncertainty reported by the proposed model reflects the uncertainty in the sparse and noisy training data, as well as the uncertainty stemming from the underlying modeling assumptions that define the SEIR-mobility epidemiology model itself (which would apply to any other epidemiology model since it is always challenging to capture the dynamics of detected patients in hundreds of different regions using one universal epidemiology model). Furthermore, the uncertainty in the estimated $R_0(t)$ is higher than the one obtained for the predictions of $C(t)$ since $R_0(t)$ is a latent variable, unlike the number of cumulative cases $C(t)$, for which observational data is available. As a consequence, it is inevitable to obtain relatively large uncertainty estimates in the predictions since the uncertainty in the model inputs (from the training data and the SEIR-mobility model limitations) is propagated to the output uncertainty, and most likely with amplification due to nonlinear interactions in the underlying dynamics.

As a consequence, the predictions of the proposed SEIR-mobility model cannot be the sole criterion for taking social and economical decisions in order to slow down the disease propagation, but can provide one additional
Fig. 10. Effects of re-opening: Real and altered workplaces mobility parameter (top left), baseline and altered estimates of $\beta(t)$ (top right), predicted effective reproduction number $R(t)$ (bottom left), and cumulative cases percentage $C(t)$ (bottom right).

4. Discussion

Amidst the COVID-19 outbreak, the modeling of the virus spread takes on life and death importance, as epidemic simulations shape national responses [41]. The reliability of such models relies heavily on their calibration on a set of incomplete and imperfect data points; a task that is often viewed as art and has been the source of scientific controversy [42]. In this work, we aspire to develop state-of-the-art machine learning algorithms to accurately determine the dynamic parameters characterizing the evolution of an epidemic, such as the time-dependent variation of the infectious rate and the effective reproduction number. Taking a systematic approach, the proposed model calibration framework will provide a quantitative understanding of the effects and efficiency of various policies designed to contain the spread of COVID-19 (and other infectious diseases) [43], and produce probabilistic forecasts for the evolution of a pandemic that can judiciously provide information for policy and decision making [44,45].

Specific to this study, the proposed approach allows us to understand the most relevant mobility and social behavior parameters that affect the spread of COVID-19. Our results suggest that working from home policies have the significant effect in containing the reproduction number. This is shown in both global and local sensitivity.
analyses, where the mobility parameter that quantifies the amount of trips to workplaces is the most relevant among all mobility and social behavior parameters. Working in confined spaces has been shown to be a risk factor for the transmission of COVID-19 [46]. On the other hand, variations in park visits showed little influence in the dynamics of pandemic, which is in line with the reduced risk of transmission while outdoors [47]. A somehow surprising result is the mild influence of the human encounters rate parameter on the reproduction number. This could be explained by the use of personal protective equipment, such as face masks, which greatly reduces the risk of transmission and by a behavior change where human encounters moved from indoor to outdoor circumstances [48]. From these results, it appears that not all human encounters lead to an increased transmission, but only those that occur indoors and for a prolonged period of time, such as in workplaces and in residential visits.

Another relevant point that our analysis revealed is the delay between changes in mobility and social behavior and changes in the reproduction number. Variations that we see today in the reproduction number are mostly influenced by changes in behavior that occurred between 17 and 21 days ago. This time delay can be explained by the incubation period of the virus [40] and by the delay in obtaining and reporting positive test results. Without any model to relate mobility and social behavior to the pandemic dynamics, policy makers are posed with a real challenge of making decisions that will have a measurable outcome only after 2 to 3 weeks. This is where tools like ours can have an impact to guide and accelerate the decisions of opening or closing certain areas. We highlight the utility of our model by simulating an increase in retail and recreation mobility in New York county, for which we observe that the reproduction number starts increasing only after 20 days of the intervention. By using the proposed model, authorities could detect the threat of increased mobility early and act in time to prevent a new outbreak.

Our methodology presents some limitations that open new research opportunities. First, there are pronounced differences in the testing frequency both in space and time. These disparities will affect the number of detected cases and bias the data that we use to train our model. Including the number of tests performed as part of the input to our model could improve our predictions and assess the effect of limited testing capabilities. However, these kinds of data are currently not available at the county level. Secondly, we have not included a potentially large fraction of cases that remain undetected or asymptomatic. This population can have an influence on the model behavior, especially when the case numbers are relatively large. Nonetheless, estimations of the undetected population based on antibody studies are currently available only at very few locations [31]. For this reason, we decided to not explicitly model this population, as we would be simply guessing its size. Finally, our model tends to perform relatively better in counties with more cases, as we can see in Fig. 4. This is intrinsically linked to the loss function that we chose to train the model. In our case, this function tends to penalize more counties with more cases. We think this behavior is desirable because it makes our predictions more accurate in regions where the pandemic is rapidly evolving. Nonetheless, it would be straightforward to adjust the loss function to emphasize other critical factors in disease transmission if needed.

CRediT authorship contribution statement

Mohamed Aziz Bhouri: Conceived the computational methods, Implemented the methods, Processed the data, Performed the simulations, Wrote the manuscript. Francisco Sahli Costabal: Conceived the computational methods, Wrote the manuscript. Hanwen Wang: Assisted with the systematic numerical studies, Wrote the manuscript. Kevin Linka: Proposed the epidemiology model and data, Wrote the manuscript. Mathias Peirlinck: Proposed the epidemiology model and data, Wrote the manuscript. Ellen Kuhl: Proposed the epidemiology model and data, Wrote the manuscript. Paris Perdikaris: Conceived the computational methods, Wrote the manuscript, Supervised all aspects of this projects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work received support from the US Department of Energy under the Advanced Scientific Computing Research program (Grant DE-SC0019116), the Defense Advanced Research Projects Agency under the Physics of Artificial Intelligence program, United States of America (grant HR00111890034), and the Air Force Office of Scientific Research, United States of America (grant FA9550-20-1-0060). M.P. and E.K. also acknowledge support from a Stanford Bio-X IIP Seed Grant, E.K. by the NIH Grant U01 HL119578, and K.L. by a DAAD Fellowship.
Appendix A. Loss function and multi-step method

The forward pass of the LSTM neural network considered in this work for one single evaluation of \( \hat{\beta}_t \) (predicted value of \( \beta \) at time instance \( t \)) is given in Algorithm 1, where \( \sigma \) refers to the sigmoid function, \( \circ \) to the element wise multiplication between two tensors of same shape and \((\cdot, \cdot)\) to the euclidean inner product between two vectors of same size. Based on the notation used in Algorithm 1, the LSTM parameters \( \theta_{NN} \) can be expressed as follows:

\[
\theta_{NN} = (W_f, W_i, W_s, W_o, U_f, U_i, U_s, U_o, b_f, b_i, b_s, b_o, V, c),
\]

where \( W_f, U_f, b_f \) refer to the forget gate’s parameters, \( W_i, U_i, b_i \) to the input/update gate’s parameters, \( W_s, U_s, b_s \) to the cell state gate’s parameters, \( W_o, U_o, b_o \) to the output gate’s parameters and \( V, c \) to the output layer parameters. \( d_h \) refers to the hidden state dimension and will be taken equal to 120. These parameters are initialized following the Xavier method [49] (see Fig. A.11).

**Algorithm 1: LSTM forward pass**

1. Inputs: \( \tau, W_f, W_i, W_s, W_o \in \mathbb{R}^{d_h \times d_h}, U_f, U_i, U_s, U_o \in \mathbb{R}^{N_m \times d_h}, b_f, b_i, b_s, b_o, V \in \mathbb{R}^{d_h}, c \in \mathbb{R}, m(t_{i-\tau+k}), k = 0, \ldots, \tau - 1 \)
2. Outputs: \( \hat{\beta}_t \)
3. \( h_{-1}, c_{-1} = 0 \)
4. for \( k = 0 : \tau - 1 \) do
5. \( f_k = \sigma[h_{k-1}W_f + m(t_{i-\tau+k})U_f + b_f] \)
6. \( i_k = \sigma[h_{k-1}W_i + m(t_{i-\tau+k})U_i + b_i] \)
7. \( c_k = \tanh[h_{k-1}W_s + m(t_{i-\tau+k})U_s + b_s] \)
8. \( o_k = f_k \circ c_{k-1} + i_k \circ \tilde{c}_k \)
9. \( h_k = o_k \circ \tanh(c_k) \)
10. end
11. \( \hat{\beta}_t = 2 \times \sigma[(h_{\tau-1}, V) + c] \)

**Fig. A.12** shows the convergence of the loss function of Eq (9) with the number of iterations. Other loss functions have been investigated, including:

1. a loss based on the residual of the differential equation of \( I(t) \) in the SEIR model satisfied by \( \hat{I}(t) \),
2. a loss based on the \( L_2 \) difference between the data of \( C(t) \) and the inferred number of cumulative cases: \( \hat{R}(t) + \hat{I}(t) \)
3. a loss based on the \( L_2 \) difference between the increase of number of cumulative cases \( C(t_i) - C(t_{i-1}) \) and the inferred number of such increase \( \hat{R}(t_i) + \hat{I}(t_i) - \hat{R}(t_{i-1}) - \hat{I}(t_{i-1}) \)
The loss defined in (9) performs better than the loss mentioned in point 1 because the latter is more prone to local minima if it is not initialized sufficiently close to the exact solution and due to the noisy data and numerical errors, the proposed loss (9) gave considerably better results. The losses in points 2 and 3 require one additional time integration to estimate the variable $R(t)$. Hence, such step introduces additional numerical errors, which explains the better performance observed with the loss defined in (9).

Various multistep methods were investigated including implicit Adams–Moulton (AM) and explicit Adams–Bashforth (AB) schemes with different orders. The implicit AM schemes perform better than the explicit AB ones thanks to their numerical stability. Higher order schemes resulted in stiffer numerical systems which sometimes led to numerical instabilities during model training. Therefore, the first order AM scheme (trapezoidal scheme) was adopted in this work.

Appendix B. Comparison against Bayesian calibration for a simple SEIR model with a constant reproduction number $\beta$

The same model described in by Eq. (1) is adopted to utilize the cumulative data. In this case, the values of $\gamma$ and $\delta$ are assumed as given and are equal to $1/6.5$ [32,33] and $1/2.5$ [11,32], as before. We assume that the parameter $\beta$ varies from one county to another, but is fixed over time and independently distributed with respect to some prior distribution. Specifically, let $\mathbf{\beta}$ be the vector where each entry represents $\beta$ of an individual county. We assume that the entries of $\mathbf{\beta}$ follow i.i.d. truncated normal distribution with mean equal to 0.5 and a scale of 0.5. Furthermore, for every county, we assume that each entry of $\log(C)$, i.e., the natural logarithm of observed cumulative case fraction, follows i.i.d. normal distributions centered at $\log(\hat{C}(\mathbf{\beta}))$ predicted by the model with $\mathbf{\beta}$, whose standard deviations follow i.i.d. log normal distributions with mean 0 and scale 10. That is,

$$\mathbf{\beta} \sim \text{TruncatedNormal}(0.5, 0.5),$$

$$\log(C) \sim \text{Normal}(\log(\hat{C}(\mathbf{\beta})), \sigma),$$

$$\sigma \sim \text{LogNormal}(0, 10).$$

Let the likelihood of the observations $\log(C)$ for a given set of parameters $\mathbf{\beta}, \sigma$ be

$$p(\log(C)|\mathbf{\beta}, \sigma),$$

$$p(\log(C)|\mathbf{\beta}, \sigma).$$
then Bayes’ theorem suggests the density of the parameters $\beta, \sigma$ given the observation $\log(C)$

$$p(\beta, \sigma | \log(C)) = \frac{p(\log(C) | \beta, \sigma)p(\beta, \sigma)}{p(\log(C))}.$$  \hspace{3cm} (B.5)

Due to the difficulty of integrating the normalization term in the denominator, the probability density of the parameter is estimated by pseudo random sampling using Hamiltonian Monte-Carlo method. Specifically, the model is calibrated via Hamiltonian Markov Chain Monte Carlo sampling using the No U-Turn Sampler (NUTS) [50] implementation of the NumPyro library [51] for 4500 iterations, with a target acceptance probability of 0.85. We then use the mean of the last 3000 samplings as an estimator for the parameter $\beta$. The relative error averaged over the extrapolation period obtained with such model is shown in the right plot of Fig. B.13 for the different counties. We also tried several different priors for $\beta$, but all of them gave similar results. In general, the simple SEIR model tends to produce cumulative cases fraction variations that are convex within the period of time considered, while the actual trajectories generally change concavity, due to the enforced behavioral intervention policies, such as lock-down and gatherings ban. Also, the sampling time is fairly long since the sampler is rejecting the candidate points because of the poor accuracy, further limiting the scalability of the method.

### Appendix C. Model sensitivity on $\gamma$ and $\delta$

Overall, the specific values chosen for $\gamma$ and $\delta$ do not have a significant impact on our model’s predictions. To support this claim, we have performed a series of systematic studies, as follows.

1. **Inferring $\gamma$ and $\delta$:** Here we extend the proposed framework so that $\gamma$ and $\delta$ are trainable constants inferred from the data. Hence, the trainable parameters $\theta$ are now given as:

$$\theta \equiv (\theta_{NN}, E^1(t_0), \ldots, E^N_c(t_0), \gamma, \delta),$$  \hspace{3cm} (C.1)

and we restrict the range of $\gamma$ and $\delta$ such that:

$$\gamma \in [1/8.09, 1/3.59] = [0.1236, 0.2786],$$

$$\delta \in [1/6.5, 1/1.5] = [0.1538, 0.6667],$$  \hspace{3cm} (C.2)

based on the range proposed by the CDC in the manual of its COVID Tracer tool [52].

We train an ensemble of 10 models with randomized parameter initializations. Once trained, the models provide estimates of the cumulative cases fraction with a relative error equal to 4.5% on the training data for the first 66 days (to be compared to 4.3% obtained with our original choice to consider $\gamma$ and $\delta$ as given), and a relative error equal to 7.7% on the testing data, corresponding to the extrapolation carried out for the remaining 15 days of the...
Fig. C.14. Convergence of the estimated $\gamma^*$ for the ensemble of 10 models with randomized parameter initializations, different lines correspond to different models.

data-set (to be compared to 6.6% obtained with $\gamma$ and $\delta$ considered as given). Therefore, the model is able to fit the data even when inferring $\gamma$ and $\delta$, and also provides acceptable extrapolation errors (only 1% higher than the extrapolation error obtained with $\gamma$ and $\delta$ considered as given). However, the identifiability of a unique parameter setting is not guaranteed. Indeed, Figs. C.14 and C.15 show the convergence of the estimated $\gamma$ and $\delta$ for the 10 models, respectively, showing that not all 10 models converge to the same final values for $\gamma$ and $\delta$. Nonetheless, the main conclusions of our sensitivity analysis study are not altered, as shown in Fig. C.16.

As a conclusion, we can assert that, thanks to the expressiveness of the LSTM network, and by modeling the susceptible to infectious transition rate $\beta$ as a function of time, the proposed framework is able to well fit the data for numerous counties and provide reliable extrapolations in time, regardless of the chosen values for $\gamma$ and $\delta$, and whether they are considered as trainable parameters or not. However, when attempting to infer $\gamma$ and $\delta$, the uniqueness of the solution is not guaranteed, which raises an issue of identifiability.

Note that the results obtained above are consistent with the fact that we only consider the cumulative cases $I + R$ as a reliable data metric. As a consequence, we cannot use the available data to trust-worthily infer the duration that a person stays in the exposed and infected compartments (which would be needed to infer $\gamma$ or $\delta$).

Also recall that the main goal of our study is to assess the influence of the different mobility and social behavior parameters on the spread of the disease. As shown in Fig. C.16, the sensitivity results are not altered by the fact of including $\gamma$ and $\delta$ within the trainable parameters, and since inferring these two parameters raises an issue of identifiability, we consider it is safer to use fixed values for such parameters based on previous existing works, and let the model infer the dependency of the susceptible to infectious transition rate $\beta$ on the mobility and social behavior parameters (along with the initial fractions of the exposed population $E(t = t_0)$ for all the different regions considered).

(2) Choosing a different fixed $\delta$: To further justify the conclusions drawn above, and to verify the robustness of the sensitivity results with respect to choosing a different values combination for $\gamma$ and $\delta$, we train 10 models with randomized parameter initializations using $\delta = 0.6$ instead of the value 0.4 used in the original manuscript, as these two values are within the range proposed by the CDC in the manual of its COVID Tracer tool [52]. Again, we verify that the sensitivity analysis results are not altered, as shown in Fig. C.17. Moreover, with $\delta = 0.6$, the models provide estimates of the cumulative cases fraction with a relative error equal to 4.3% on the training data for
the first 66 days (exactly equal to what we obtained with $\delta = 0.4$), and a relative error equal to 7.9% on the testing data, corresponding to the extrapolation carried out for the remaining 15 days of the data-set (to be compared to 6.6% with $delta = 0.4$).

Overall, we can state that $\delta$ and $\gamma$ parameters are disease-specific and we fixed their values based on the information available at the time we carried out this study. The average of the values obtained in the works that we cited corresponds to $\delta = 1/2.5$ and $\gamma = 1/6.5$. There are also some reasons for why one may consider different values:

1. More information and data are being collected and reported every single day,
Fig. C.17. $s_{i,j}$: Sensitivity of susceptible to infectious transition rate $\beta$ to the mobility and social behavior parameters with $\delta = 0.6$.

2. but $\gamma$ can also change based on
   - mutations of the virus and thus different disease time courses with the virus mutation that was initially going through the population
   - increased insights into how to treat the disease and suppress the disease faster now than at the start when not much was known.

However, changes in the $\gamma$ and $\delta$ values due to the reasons above are not expected to impact the $\beta(t)$ time evolution that our model learns because we only work with the confirmed cases which include all potential virus mutation types, and the number of confirmed cases is independent of the disease treatment. This is confirmed by the sensitivity results given above, indicating that $\delta$’s variation does not really alter the results unless there is a significant change, which is not the case if we refer to the ranges given by the CDC.

Appendix D. Predicted forecasts for a 21-day window

The 15 day-prediction window was chosen since it is typical to most of the existing studies [53,54]. Moreover, for the considered counties, the corresponding data-set has 81 days, which only leaves 66 days for the training data-set if a 15-day prediction is adopted.

Nonetheless, additional simulations are performed in order to assess the robustness of the results when using training data from the first 60 days, while the remaining 21 days are considered for extrapolation as testing data. We train an ensemble of 10 models with randomized parameter initializations. Once trained, the models provide estimates of the cumulative cases fraction with a relative error equal to 4.9% on the training data for the first 60 days (to be compared to 4.3% obtained with a training data-set of 66 days), and a relative error equal to 11.5% on the testing data, corresponding to the extrapolation carried out for the remaining 21 days of the data-set (considerably higher than the 6.6% obtained with a testing data-set of 15 days, as expected).

Furthermore, as it can be seen in comparison to the 15-day predictions, the uncertainty in the estimated $R_0(t)$ grows for some counties (since $R_0(t)$ is a latent variable compared to the tighter uncertainty obtained for $C(t)$). Aiming for a more extended prediction range would inevitably result in a higher uncertainty, even for $C(t)$ as shown in Fig. D.18.

We also verify that the main conclusions of our sensitivity analysis study are not altered when considering the training data-set of 60 days, as shown in Fig. D.19.

Finally, most COVID-19-related political decisions are made based on two week-predictions [55], which justifies the 15 days choice. Such interval is actually chosen as the upper limit of the virus incubation period [56] since political decisions taken on a given day would have an impact on the virus spread after a period of around 2 weeks.
Therefore we believe that our initial choice of a 15-day prediction window is reasonable, and we are optimistic that it can be stretched to accommodate longer time forecasts if more data is used for training (as the latter becomes available with the evolution of the pandemic).
Fig. D.19. $s_{i,j}$: Sensitivity of susceptible to infectious transition rate $\beta$ to the mobility and social behavior parameters with models using a training dataset of 60 days.

References


