Asymptotic estimates of SARS-CoV-2 infection counts and their sensitivity to stochastic perturbation

Davide Faranda, Isaac Pérez Castillo, Oliver Hulme, Aglaé Jézéquel, Jeroen Lamb, Erica Thompson

Asymptotic estimates of SARS-CoV-2 infection counts and their sensitivity to stochastic perturbation

Davide Faranda*
Laboratoire des Sciences du Climat et de l'Environnement,
CEA Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ,
Université Paris-Saclay & IPSL, 91191 Gif-sur-Yvette, France
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK and
LMD/IPSL, Ecole Normale Superieure,
PSL research University, Paris, France

Isaac Pérez Castillo
Department of Quantum Physics and Photonics, Institute of Physics,
UNAM, P.O. Box 20-364, 01000 Mexico City, Mexico and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK

Oliver Hulme
Danish Research Centre for Magnetic Resonance,
Centre for Functional and Diagnostic Imaging and Research,
Copenhagen University Hospital Hvidovre,
Kettegard Allé 30, 2650, Hvidovre, Denmark and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK

Aglaé Jezequel
LMD/IPSL, ENS, PSL Université, École Polytechnique,
Institut Polytechnique de Paris, Sorbonne Université, CNRS, Paris France and
Ecole des Ponts, Marne-la-Vallée, France

Jeroen S.W. Lamb
Department of Mathematics, Imperial College London, UK and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK

Yuzuru Sato

RIES/Department of Mathemarics, Hokkaido University,
N20 W10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK

Erica L. Thompson
Centre for the Analysis of Time Series,
London School of Economics and Political Science,
Houghton Street, London WC2A 2AE and
London Mathematical Laboratory, 8 Margravine Gardens, London, W6 8RH, UK

(Dated: March 25, 2020)

Abstract

Despite the importance of having robust estimates of the time-asymptotic total number of infections, early estimates of COVID-19 show enormous fluctuations. Using COVID-19 data for different countries, we show that predictions are extremely sensitive to the reporting protocol and crucially depend on the last available data-point, before the maximum number of daily infections is reached. We propose a physical explanation for this sensitivity, using a Susceptible-Exposed-Infected-Recovered (SEIR) model where the parameters are stochastically perturbed to simulate the difficulty in detecting asymptomatic patients, different confinement measures taken by different countries, as well as changes in the virus characteristics. Our results suggest that there are physical and statistical reasons to assign low confidence to statistical and dynamical fits, despite their apparently good statistical scores. These considerations are general and can be applied to other epidemics.

I. LEAD PARAGRAPH

COVID-19 is currently affecting over 180 countries in the world and poses serious threats to public health as well as economic and social stability of many countries. Modeling and extrapolating in near real-time the evolution of COVID-19 epidemics is a scientific challenge, which requires a deep understanding of the non-linearities undermining the dynamics of the epidemics. Here we show that real-time predictions of COVID-19 infections are extremely sensitive to errors in data collection and crucially depend on the last available data-point. We test these ideas in both statistical (logistic) and dynamical (Susceptible-Exposed-Infected-Recovered) models that are currently used to forecast the evolution of the COVID-19 epidemic. Our goal is to show how uncertainties arising from both poor data quality and inadequate estimations of model parameters (incubation, infection and recovery rates) propagate to long term extrapolations of infections count. We provide guidelines for reporting those uncertainties to the scientific community and the general public.

* Correspondence to davide.faranda@lsce.ipsl.fr

II. INTRODUCTION

SARS-CoV-2, a zoonotic virus of the coronavirus family [1], that provokes an infectious disease known as COVID-19, has emerged in China at the end of 2019, affecting first the Hubei province and quickly spreading to all Chinese provinces [2]. The failure of initial containment measures caused the virus to spread internationally, and on March 11th, The World Health Organization (WHO) declared COVID-19 a pandemic [3]. According to the WHO Situation Report-59 released on March 19th [4], the number of countries affected by the pandemic is 176, with 209 839 confirmed infections and 8778 deaths. As this report also notices: *the number of confirmed cases worldwide has exceeded 200 000. It took over three months to reach the first 100 000 confirmed cases, and only 12 days to reach the next 100 000*, an astonishing development, due to the highly contagious character of SARS-CoV-2.

SARS-CoV-2 causes potentially life-threatening form of pneumonia in a non-negligible patients fraction [5]. Enormous efforts to contain the virus and to not overwhelm intensive care facilities are currently taken all over the world. Following the drop in infections observed in the Hubei province, restrictive confinement measures have been taken in many countries [6]. Most of the time, those measures are taken on the basis of epidemics models, which are fitted with dynamical or statistical models on the available data.

COVID-19 data should be provided daily, following a request of the WHO. To date, the WHO guidelines require countries to report, at each day t, the total number of infected patients $I(t)$ as well as the number of deaths $D(t)$. Unfortunately, there is large variability in the way both $I(t)$ and $D(t)$ are counted. We provide some illustrative example: on the one hand, Italy shows the highest fatality rate:

$$f = \sum_{t=1}^{\tau} D(t) / \sum_{t=1}^{\tau} I(t) \simeq 0.07 \tag{1}$$

possibly because $D(t)$ includes all deaths who have contracted SARS-CoV-2, independently on whether the virus is the first cause of death. Moreover, in a recent interview [7], Italian biologist Bucci has stated that $D(t)$ can be underestimated because it does not include those patients who died at home without being tested. On the other hand, in Germany, the fatality rate is extremely low $f \simeq 0.002$. There may be several explanations

for this : some query data methodology (e.g. a different method to determine $D(t)$) while others say high testing rates are giving a more accurate picture [8].

Great uncertainties also exist in the count of $I(t)$. Whereas in the early stage of the epidemic several countries tested asymptomatic individuals to track back the infection chain, recent policies to estimate $I(t)$ have changed. Most of the western countries now test only patients displaying severe SARS-CoV-2 symptoms. In an effort of tracking all the chain of infections, South Korea has tested many asymptomatic people. This latter strategy has proven effective in supporting actions to reduce the rate of new infections. A recent study [9] has estimated that an enormous part of total infections were undocumented (80% to 90%) and that those undetected infections were the source for 79% of documented cases in China.

The goal of this paper is to analyse the effect of those large uncertainties in real-time forecasting of the long term behavior of the COVID-19 epidemic [10]. As stated by Polonsky et al [11], there is a need for defining robust methods to assess both the intrinsic errors inherent to fitting procedures as well as those introduced by poor data-quality. Funk et al [12] give a concrete example of this applied to the Ebola epidemics in the Western Area region of Sierra Leone in 2014-15. Classically, epidemiologists rely on Susceptible-Exposed-Infected-Recovered (SEIR) models [13]. These models consist of ordinary differential equations where a population is divided into compartments, with the assumption that every individual in the same compartment has the same characteristics. In SEIR, population is divided into Susceptible, Exposed, Infected and Recovered individuals. Such models predict a sigmoid shape of the total number of infections $C(t) = \sum_{t=1}^{\tau} I(t)$. Using the available national data points $I(t)$ one can obtain long term estimates on the total of COVID-19 infections in each country. This paper focuses on the estimation of the sensitivity of these models to the last available data point, before the inflection point of the $I(t)$ curve is reached. We use SEIR models to show the possible origins of this sensitivity by perturbing the relevant parameters, often assumed deterministic, with a noise that mimics changes in the way the virus is spreading, e.g. as a result of application of confinement measures, or the presence (rate/magnitude) of super-spreaders [14]. The paper is organised as follows: in Section III we discuss the various sources of data for COVID-19 and their shortcomings, and then we discuss in detail the SEIR model and its statistical modelling. In Section IV we discuss the

results focusing on the statistical sensitivity of the modelling, and apply it to data from France, UK and Italy. We finish, in Section V, with some remarks and point out some potentially beneficial policy guidelines.

III. DATA AND MODELLING

A. Data

The data repository used in this paper for COVID-19 data is a Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data repository [15] is also supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). We used datasets of cases confirmed with a laboratory test, irrespective of clinical signs and symptoms [3]. The data contains, as recognized by the public authorities that dispatched them, several inhomogeneities due to the different ways of testing patients with suspicious symptoms. As an example, Italy announced on Feb. 26 that it relaxed testing criteria to the point that contacts linked to confirmed cases or recent travelers to outbreak areas would not be tested anymore, unless they show symptoms. Unlike Italy, South Korea (population of 51 million) is testing 15000 to 20000 individuals per day since Feb. 27 with the goal to minimize hospital pressure and stop the epidemics in the early stages [16]. COVID-19 data also suffers from reporting problems due to the local management of health infrastructures. In Italy, healthcare is a regional task and everyday data are collected at a regional level and transmitted to the Protezione Civile, who transfers the data to WHO. Many inconsistencies and delays have been documented in this transfer process [17]. A similar situation occurs in Mexico, in which for instance, private institutions, either hospitals or laboratories, do not possess the necessary national and international certifications given by the *Instituto de Diagnóstico y Referencia Epidemiológicos* (InDRE) and therefore their tests are not considered valid and must be redone by certified institutions[18], thus unnecessarily delaying the release of accurate daily reports. COVID-19 data of Mexico was collected from the daily reports generated by Mexico's *Secretaría de Salud*[19]. Our goal is to account for these uncertainties in the modelling of COVID-19 data.

B. An epidemiological Susceptible-Exposed-Infected-Recovered model

The Susceptible-Exposed-Infected-Recovered (SEIR) model [13] is an epidemiological compartmental model where a total population N is divided into susceptible individuals S, exposed individuals E, infected individuals I, and the number R of people who have had the disease and are now either recovered or dead (and assumed not to be susceptible to reinfection). The model is constructed under the assumption that the total population $N = S(t) + E(t) + I(t) + R(t)$ does not vary. This implies:

$$0 = dN/dt = dS/dt + dE/dt + dI/dt + dR/dt, \quad \forall t \geq 0. \tag{2}$$

The model relies on some assumptions. First of all, susceptible individuals end up becoming infected and infected individuals can only recover or die. Individuals who are exposed (E) have had contact with an infected person, but are not themselves infectious. Furthermore, those who have recovered or died are forever immune. It is also assumed that susceptibility is equal for all and that it is proportional to the product of $I(t)$ and $S(t)$ at a time t. These assumptions lead us to a set of four ordinary differential equations:

$$\frac{dS}{dt} = -\lambda S(t)I(t) \tag{3}$$

$$\frac{dE}{dt} = \lambda S(t)I(t) - \alpha E(t) \tag{4}$$

$$\frac{dI}{dt} = \alpha E(t) - \gamma I(t) \tag{5}$$

$$\frac{dR}{dt} = \gamma I(t). \tag{6}$$

Here $\gamma > 0$ represents the mean recovery/death rate, or $1/\gamma$ the mean infection period, $\lambda = \lambda_0/S(0) > 0$ is considered the contact or infection rate of the disease and it is rescaled by the initial number of susceptible individuals $S(0)$ and α is the inverse of the incubation period. These expressions satisfy (2) as required. Because data are reported only on a daily basis, we adopt the discrete SEIR model:

$$S(t+1) = S(t) - \lambda S(t)I(t) \tag{7}$$

$$E(t+1) = (1-\alpha)E(t) + \lambda S(t)I(t) \tag{8}$$

$$I(t+1) = (1-\gamma)I(t) + \alpha E(t) \tag{9}$$

$$R(t+1) = R(t) + \gamma I(t). \tag{10}$$

This model is obtained rewriting the ordinary differential equations 3-6 with an Euler Scheme and fixing $dt = 1$ day. An important derived quantity of the model is $R_0 = \lambda_0/\gamma$, the average reproduction number of the virus in a population. This quantity represents the number of cases, on average, an infected person will cause during their infectious period. For COVID-19 in Wuhan in January 2020, $R_0 = 2.68$ with 95% CrI 2·47–2·86 according to an estimate performed with Wuhan data [20]. Dynamical modelling of COVID-19 epidemic has been proposed in [21]. In that study, the authors used a Susceptible-Exposed-Infected-Recovered model with delays and performed a sensitivity study on the parameters. Fixing $\lambda \simeq 1$ as in [21] and $\gamma = 0.37$ to recover the value of R_0 found in [20] (assuming that the behavioural elements of viral transmission are consistent in other populations), we are left with the choice of α. The range for incubation period of SARS-CoV-2 has been determined in [22] between 2 and 11 days. As a comparison, this range is estimated to be between 2 and 5 days for human coronavirus, and between 2 and 10 days for severe acute respiratory syndrome (SARS) coronavirus [23]. Using a trial and error procedure and a subjective estimation of the quality of the fit, we obtain the best fit when we set $\alpha = 0.27$ (corresponding to an incubation period between 3 and 4 days) and initial conditions $S(0) = 33000$, $I(0) = 2$, $E(0) = R(0) = 0$. The fit against the Chinese data is reported in Figure 1. We are aware that a log-likelihood method with cross-validation would provide a better fit as well as an estimate of the uncertainty. However, we underline that the goal of this work is not to provide the best possible model but rather to explore the sensitivity of it to perturbations. First of all, we note that, despite its simplicity, the model shows qualitatively similar behaviour to the published data. Note that there is a discontinuity in the dataset, which is due to a change in the way infections were counted, introduced on Feb. 12, 2020 [24].

This model has also evident deficiencies in representing the COVID-19 infections. First of all the total population N here is to be intended as a number of people who could have been in contact with infected individuals. Furthermore, the population under consideration does not consist of a group of about the same age and general health level, and the the group members do not mix homogeneously. The model does not have any spatial component, nor does it predict the influences of policy and behavioural responses to the progress of the pandemic. More complex models introducing further parameters would likely lead to overfitting and over-confident predictions, due to the limited volume of data currently available. No model will be sufficient to predict the outcome of this pandemic: the outcome depends on our

response. Models are presented here with the aim of generating some insight into the overall behaviour and the risks entailed by inaction.

C. Statistical Modelling

When insight is limited and compartmental models are not suited, phenomenological statistical models provide a starting point for estimation of key transmission parameters, such as the reproduction number, and forecasts of epidemic impact [25]. One of the simplest ways to model the epidemics is to observe that the function $C(t)$ is a sigmoid function and perform a statistical fit of the data to extrapolate the long-term behavior of the epidemics [26, 27]. Among all the possible sigmoid functions, two have proven useful in fitting epidemic growth: the generalized logistic distribution [28] and the generalized Gompertz distribution [29]. A complete overview of sigmoid functions is presented in [30], although applied to in a different context. Since our considerations are independent of the sigmoid function used, we will present results for the generalized logistic model only. The model reads:

$$C(t) = a/(1 + b \cdot \exp(-c \cdot t)); \tag{11}$$

where a, b and c are parameters of the model. They are linked in a non-explicit way to the solution of the SEIR model. A fit to the Chinese data is presented in Figure 2. Logistic fits are performed with the MATLAB Nonlinear least-squares solver constraining objective function with gradient. At first sight, one can be tempted to use $R^2 \simeq 0.997$ as a quality indicator of the fit. However, we stress that R^2 is not an appropriate measure for nonlinear regression models: given the smoothness of data, there will be lots of models (eg low-order polynomial) which could fit well (get a very good R^2) but would not make credible predictions [31]. These data are however collected at a mature stage of the epidemic and as such the characteristics of the logistic fit to these data can be assigned with greater confidence. In the next section we will discuss the performance of the statistical model in the early stage of the epidemics, where the logistic function can be used to extrapolate the behavior of $C(t)$.

IV. RESULTS: STATISTICAL AND DYNAMICAL MODELLING OF EARLY STAGES OF THE EPIDEMICS

A. Statistical sensitivity

We begin by showing the sensitivity of the logistic extrapolations in the early stage of the epidemics by looking at French data from Mar. 04 to Mar 20. France has previously recorded sporadic cases of SARS-CoV-2 infections but the exponential growth phase started at the beginning of March 2020. To show the high sensitivity to the last point of the datasets we first perform a logistic fit with data starting from different dates and ending Mar. 20 (Figure 3a) and then do the reverse experiment by fitting data starting on Mar. 04 but ending at different dates (Figure 3b). Clearly, fits are more stable by removing days from the beginning of the outbreak than from the most recent past. Again, we stress the inadequacy of the R^2 metric as it yields values above $R^2 > 0.99$ for all cases considered in Figure 3. The analysis suggests that, if a large error is presented in the last data point, the extrapolation has less predictive adequacy. This implies very narrow estimates of confidence intervals for $C(t)$: for each fit, confidence intervals are as small as the thickness of the line used in the plots in Figure 3. This prevents a correct evaluation of the confidence interval, which is critical to assess the uncertainties around the future evolution of the epidemics, and to build relevant policies to address the worst case scenario.

To further test this concept, we now assume we are uncertain about the magnitude of the last data point $C(t^*)$. To simulate this uncertainty, we replace it with a random number $\xi(t^*)$ drawn from a discrete uniform distribution with mean $C(t^*)$ and standard deviation $0.2C(t^*)$. We therefore construct an ensemble of 100 possible trajectories under this generative process. Results are presented in Figure 4 for UK (a), France (b) and Italy (c). To date, Italy is at a more mature stage of the epidemic, while France and UK face an earlier stage. This is reflected in the spread of the ensemble: for the UK, forecasting the epidemic with a logistic fit is not informative of the course of the epidemic: the ensemble spread just suggests that the current phase is an exponential growth and at best it can inform that worst case scenarios should be considered at this point. The ensemble spread reduces when the epidemics is at a more mature stage (Italy). Indeed, if we set $b = 1$ and

we start the fit from time t_0 then the logistic distribution is written:

$$C(t) = a/(1 + \exp(-c(t - t_0))).$$

In the early growth phase, $\exp(-c(t - t_0)) \gg 1$, so:

$$C(t) \sim a \exp(c(t - t_0)) = a \exp(-c \cdot t_0) \exp(-c \cdot t) = A \exp(-c \cdot t)$$

.

Even though we can fit A and b to data, recalling that $A = a \exp(-c \cdot t_0)$ we have huge freedom over a, the upper asymptote that determines the final count of the epidemics.

B. Dynamical sensitivity in a stochastic SEIR model

Another way to understand the sensitivity in epidemics is to release the assumption that incubation period α, infection rate λ and recovery rate γ are constant through the epidemics [32]. Intrinsically they can vary, because of the presence of individuals with an extremely high transmission rate known as super-spreaders [14], or due to the release or the application of confinement measures, or changes in the SARS-CoV-2 characteristics. They can also display spurious variations due to the way data are reported or collected, for the problems specified above. We explore all these possibilities by considering α, λ, and γ as time varying processes. The idea of using stochastic models to represent epidemics is not new to the literature [33–35]. In the modelling of COVID-19 infections can be further justified by the evidence that $R_0 = \lambda/\gamma$ displays spatial and temporal variability [11]. For example, Wu et al. [20] show fluctuations of R_0 in different Chinese regions. These differences are due to changes in the duration of contagiousness, likelihood of infection per contact and the contact rate [36] which depends on demographic spatial variability [37]. There is however little consensus on which variables or parameters should be perturbed in order to get a realistic behavior. Our goal here is different than obtaining the best possible forecasts of the epidemics as we want to understand which parameter causes a large sensitivity in the final $C(t)$ counts. Let us begin, by alternatively replacing in Equations 7-10 one of the constant parameters $\kappa \in \{\alpha, \lambda, \gamma\}$ with a stochastic process:

$$\kappa(t) = \kappa_0 + \sigma \cdot \xi(t) \tag{12}$$

where σ is the intensity of the perturbation and $\xi(t)$ a random variable drawn from a normal distribution $N(0,1)$ at each time. The purpose of equation 12 is to introduce instantaneous discrete jumps in the values of the daily parameters. This discrete process, used in [38], is more appropriate than a continuous one (see, e.g. [39]) when observations are affected by large detection errors, as in the present case. Figure 5 shows an example of 30 realisations of a stochastic SEIR COVID-19 model, obtained by replacing alternately α (a,b), λ (b,d) and γ with the stochastic process in Eq 12 and using $\sigma = 0.2\kappa_0$ to get fluctuations of the order of 20% of each parameter values, in analogy with the statistical sensitivity studies performed in the previous section. The sensitivity clearly depends on the perturbed parameter: a perturbation on α mostly implies a different timing of the epidemics while the final cumulative number of infections $C(t)$ remains unchanged. Perturbations on λ and γ affect the final $C(t)$ in a deeper way, leading to a total variation in the number of cases of the order of 20%. Indeed, by changing λ and γ, we also modify the basic reproduction number R_0. The idea of having a time-varying reproduction number has been already exploited in [40], although the authors have directly modelled the dynamics of a dynamic reproduction number $R(t)$ without introducing a SEIR model.

As a further step, we add noise simultaneously to all parameters of the SEIR model via Equation 12. Six realisations of the model are shown in Figure 6. Figure 6-a,b) shows the evolution of $S(t)$, $R(t), E(t)$ and $C(t)$. We have separated the time evolution of $I(t)$ in Figure 6-c) to compare it with that of COVID-19 data for China, South Korea and Italy (Figure 6d). Despite having a quasi-smooth behavior of $C(t)$, we observe a highly non-smoothness of $I(t)$, which is reflected by the data. The sensitivity of the model is higher when $I(t)$ is large, because γ and λ directly act on $I(t)$. Therefore, when approaching the maximum of $I(t)$ ($t \sim 50$ days) small changes in the parameters can greatly affect the final total count of infections $C(t)$. This implies that mitigation strategies based on the reduction of λ by self-isolation, social distancing, are way more effective if imposed at the early stage of the epidemics, as they can suppress positive fluctuations of and help reducing R_0.

V. DISCUSSION

In this work we have discussed the statistical and dynamical sensitivity of asymptotic estimates of COVID-19 infections when performed at the early stages of the epidemics. First of all, we noted that SEIR model, with λ, γ and α inferred from clinical studies, can fit Chinese data with a value of $N \simeq 33000$ that is very different from that of the Chinese, Hubei or Wuhan populations. This enormous discrepancy can be due both to a large underestimation in the number of total cases, or to the effectiveness of confinement measures which results in a smaller exposed population. This estimate should be taken as a first caveat in fitting a SEIR model to infer COVID-19 epidemics evolution in other countries as results may be largely under/over-estimated [11].

Then, we have shown that statistical fits often used to extrapolate the long term behavior of the epidemics are greatly affected by the magnitude of the last data point, despite values of R^2 close to one, leading to unrealistic or over-confident estimates of confidence intervals on the forecast of the total number of infections [41, 42]. In the early stage of the epidemics, we have shown that knowing the last data point with a relative 20% error, can lead to a final extrapolation of infections with an error of several orders of magnitude. In order to improve the estimates of statistical models one should replace R^2 estimates by a formal comparison of model-alternatives using information criteria (e.g. AIC or BIC) or a log-likelihood approach with a leave-one-out cross-validation procedure. A simple cross validation can follow both the approaches described in this paper: i) exclude the last data points and check the stability of the estimates, ii) add noise to the last data point and obtain an ensemble of estimates. Another approach could be based on evaluating every day each model on the performance in predicting the new data point, and then used again with the new data point for an updated estimate.

Finally, we have investigated whether this statistical sensitivity can be dynamically reproduced with a SEIR model where parameters are considered stochastic processes (Equation 12). We have found that the stochastic dynamics are more sensitive to γ and λ. Perturbations on these parameters are proportional to the number of infected patients $I(t)$ and are therefore important in the growth phase of the epidemics. Actual data display fluc-

tuations even larger than those simulated in the stochastic models, suggesting that instead of assuming observational Gaussian noise on the parameters, jump processes (e.g. Levy noise) may be more appropriate [43]. Furthermore, we noticed that large fluctuations in the number of detected infections is also due to changes in the testing protocols and availability of tests. All these inconsistencies prevent the possibility of performing meaningful asymptotic statistical or dynamical modelling for COVID-19, or comparing results among different countries. This may be even more problematic in least developed countries, which are just beginning to register cases [44–46].

Our study suggests that dynamical and statistical modelling should focus on limited stages of the epidemics and restrict the analysis to specific regions, accounting for large uncertainties as done in [47]. Modelling approaches should take into account both statistical uncertainties as well as expert knowledge in a sort of Bayesian framework that allows to guide the choice of prior probabilities [10]. In the interest of preserving the public health of as many individuals as possible, once modelled the uncertainty in the data, the worst case scenarios should always be taken into account very seriously as a guideline to enforce strict confinement measures.

VI. ACKNOWLEDGMENTS

This paper is dedicated to the memory of F Molinari, who recently passed away from COVID-19. DF acknowledges A Adamou, B Dubrulle, F Pons, F Daviaud, P Yiou, M Kagayema, S Fromang and G Ramstein for useful discussions.

VII. DATA AVAILABILITY

The data that support the findings of this study are openly available in `https://systems.jhu.edu/research/public-health/ncov/`, maintained by Johns Hopkins Uni-

versity Center for Systems Science.

[1] Eleanor R Gaunt, Andrew Hardie, Eric CJ Claas, Peter Simmonds, and Kate E Templeton. Epidemiology and clinical presentations of the four human coronaviruses 229e, hku1, nl63, and oc43 detected over 3 years using a novel multiplex real-time pcr method. *Journal of clinical microbiology*, 48(8):2940–2947, 2010.

[2] Jin Wu, Weiyi Cai, Derek Watkins, and James Glanz. How the virus got out. *The New York Times*.

[3] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 51. 2020.

[4] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 59. 2020.

[5] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 395(10223):497–506, 2020.

[6] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet*, 395(10228):931–934, 2020.

[7] Luca Fraioli. Bucci: "dalla lombardia numeri ormai insensati. i contagiati sono di più". *Repubblica*, Mar 2020.

[8] Philip Oltermann. Germany's low coronavirus mortality rate intrigues experts. *The Guardian*.

[9] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020.

[10] Angel N Desai, Moritz UG Kraemer, Sangeeta Bhatia, Anne Cori, Pierre Nouvellet, Mark Herringer, Emily L Cohn, Malwina Carrion, John S Brownstein, Lawrence C Madoff, et al. Real-time epidemic forecasting: Challenges and opportunities. *Health security*, 17(4):268–275, 2019.

[11] Jonathan A Polonsky, Amrish Baidjoe, Zhian N Kamvar, Anne Cori, Kara Durski, W John Edmunds, Rosalind M Eggo, Sebastian Funk, Laurent Kaiser, Patrick Keating, et al. Out-

break analytics: a developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B*, 374(1776):20180276, 2019.

[12] Sebastian Funk, Anton Camacho, Adam J Kucharski, Rachel Lowe, Rosalind M Eggo, and W John Edmunds. Assessing the performance of real-time epidemic forecasts: A case study of ebola in the western area region of sierra leone, 2014-15. *PLoS computational biology*, 15(2):e1006785, 2019.

[13] Fred Brauer. Compartmental models in epidemiology. In *Mathematical epidemiology*, pages 19–79. Springer, 2008.

[14] James O Lloyd-Smith, Sebastian J Schreiber, P Ekkehard Kopp, and Wayne M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

[15] Data last downloaded on Mar. 23 from @x `https://systems.jhu.edu/research/public-health/ncov/`.

[16] Kim Arin. Drive-thru clinics, drones: Korea's new weapons in virus fight. *The Korea Herald*.

[17] Pagella Politica AGI. Come vanno letti i dati sul coronavirus in italia. *AGI Agenzia Italia*.

[18] As of 20th of March 2020 only two private hospitals in Mexico have been certified by InDRE to carry out tests.

[19] There is a delay between the data reported daily by the WHO and that reported by Mexico health authorities.

[20] Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 395(10225):689–697, 2020.

[21] Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic analysis of covid-19 in china by dynamical modeling. *arXiv preprint arXiv:2002.06563*, 2020.

[22] 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*, 2020.

[23] Justin Lessler, Nicholas G Reich, Ron Brookmeyer, Trish M Perl, Kenrad E Nelson, and Derek AT Cummings. Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet infectious diseases*, 9(5):291–300, 2009.

[24] Amy Gunia and Michael Zennie. China reported a huge increase in new covid-19 cases. here's why it's actually a step in the right direction. *Time*.

[25] Gerardo Chowell, Doracelly Hincapie-Palacio, Juan Ospina, Bruce Pell, Amna Tariq, Sushma Dahal, Seyed Moghadas, Alexandra Smirnova, Lone Simonsen, and Cécile Viboud. Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics. *PLoS currents*, 8, 2016.

[26] Gerardo Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, 2(3):379–398, 2017.

[27] Raimund Bürger, Gerardo Chowell, and Leidy Yissedt Lara-Díiaz. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. *Mathematical biosciences and engineering: MBE*, 16(5):4250–4273, 2019.

[28] Pierre-François Verhulst. Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.*, 10:113–126, 1838.

[29] Benjamin Gompertz. Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. *Philosophical transactions of the Royal Society of London*, (115):513–583, 1825.

[30] IJ Wellock, GC Emmans, and I Kyriazakis. Describing and predicting potential growth in the pig. *Animal Science*, 78(3):379–388, 2004.

[31] Andrej-Nikolai Spiess and Natalie Neumeyer. An evaluation of r 2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a monte carlo approach. *BMC pharmacology*, 10(1):6, 2010.

[32] Hao Xiong and Huili Yan. Simulating the infected population and spread trend of 2019-ncov under different policy by eir model. *Available at SSRN 3537083*, 2020.

[33] Lars Folke Olsen and William Morris Schaffer. Chaos versus noisy periodicity: alternative hypotheses for childhood epidemics. *Science*, 249(4968):499–504, 1990.

[34] Hakan Andersson and Tom Britton. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

[35] Joseph Dureau, Konstantinos Kalogeropoulos, and Marc Baguelin. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555, 2013.

[36] Giulio Viceconte and Nicola Petrosillo. Covid-19 r0: Magic number or conundrum? *Infectious Disease Reports*, 12(1), 2020.

[37] Ilya Kashnitsky. Covid-19 in unequally ageing european regions. 2020.

[38] Davide Faranda and Sandro Vaienti. Extreme value laws for dynamical systems under observational noise. *Physica D: Nonlinear Phenomena*, 280:86–94, 2014.

[39] Davide Faranda, Yuzuru Sato, Brice Saint-Michel, Cecile Wiertel, Vincent Padilla, Bérengère Dubrulle, and François Daviaud. Stochastic chaos in a turbulent swirling flow. *Physical review letters*, 119(1):014502, 2017.

[40] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020.

[41] Andrea Remuzzi and Giuseppe Remuzzi. Covid-19 and italy: what next? *The Lancet*, 2020.

[42] Choujun Zhan, K Tse Chi, Zhikang Lai, Tianyong Hao, and Jingjing Su. Prediction of covid-19 spreading profiles in south korea, italy and iran by data-driven coding. *medRxiv*, 2020.

[43] Xianghua Zhang and Ke Wang. Stochastic seir model with jumps. *Applied Mathematics and Computation*, 239:133–143, 2014.

[44] Joost Hopman, Benedetta Allegranzi, and Shaheen Mehtar. Managing covid-19 in low-and middle-income countries. *JAMA*, 2020.

[45] Marius Gilbert, Giulia Pullano, Francesco Pinotti, Eugenio Valdano, Chiara Poletto, Pierre-Yves Boëlle, Eric D'Ortenzio, Yazdan Yazdanpanah, Serge Paul Eholie, Mathias Altmann, et al. Preparedness and vulnerability of african countries against importations of covid-19: a modelling study. *The Lancet*, 2020.

[46] J Steenhuysen and S Nebehay. Countries rush to build diagnostic capacity as coronavirus spreads. *Reuters*, 2020.

[47] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, and Chris P Jewell. Novel coronavirus 2019-ncov: early estimation of epidemiological parameters and epidemic predictions. *MedRxiv*, 2020.

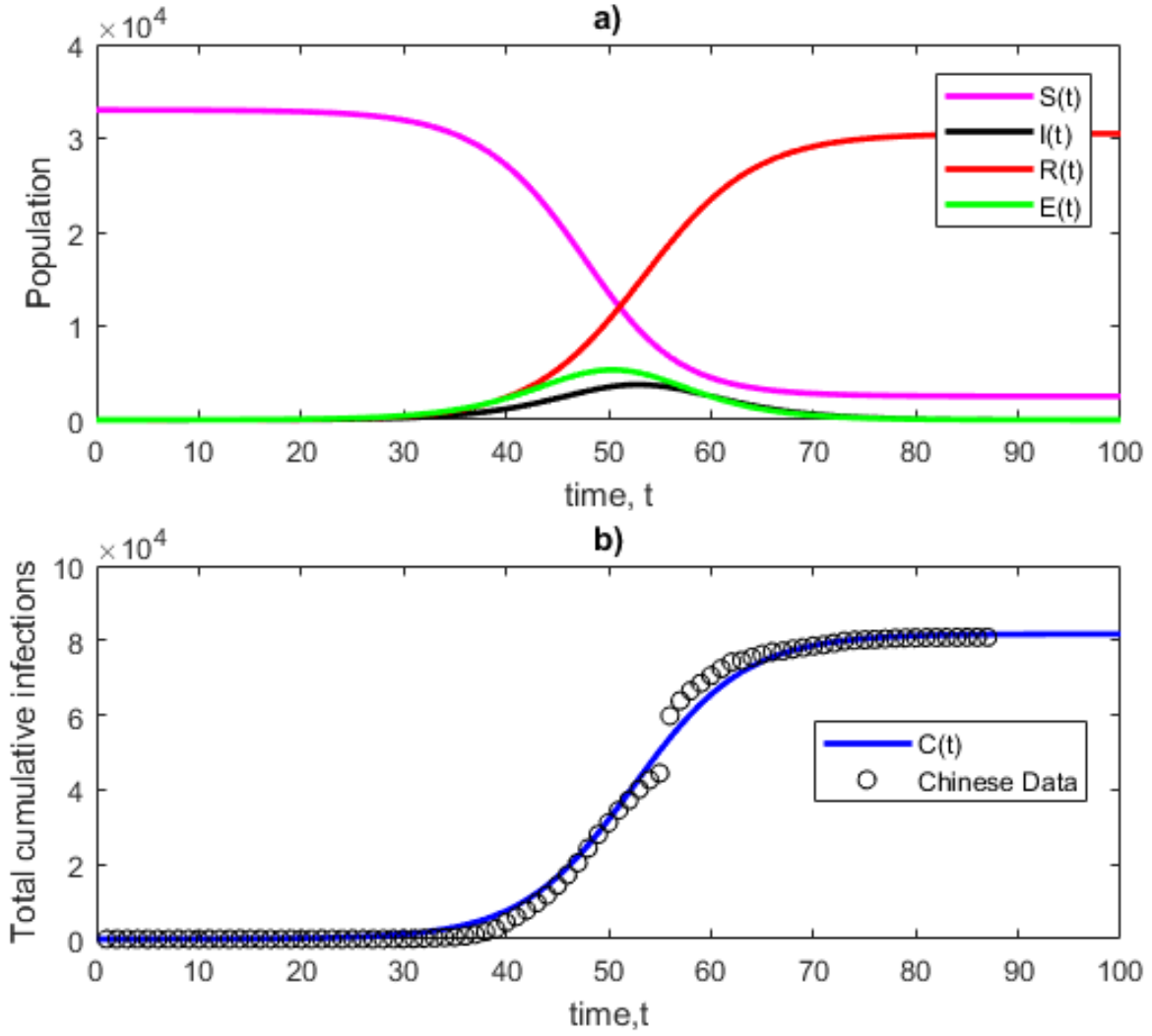

FIG. 1. Example of a Susceptible-Exposed-Infected-Recovered (SEIR) model of COVID-19 (Eqs 7-10) with $\lambda = 1./S(0)$, $\alpha = 0.27$, $\gamma = 0.37$. Initial conditions are set to $I(0) = 2$, $S(0) = 33000$, $E(0) = R(0) = 0$. a) Time evolution for the variables of the system, b) Time evolution for the total number of infections $C(t)$ against the Chinese data with t=1 corresponding to Dec 19. 2019.

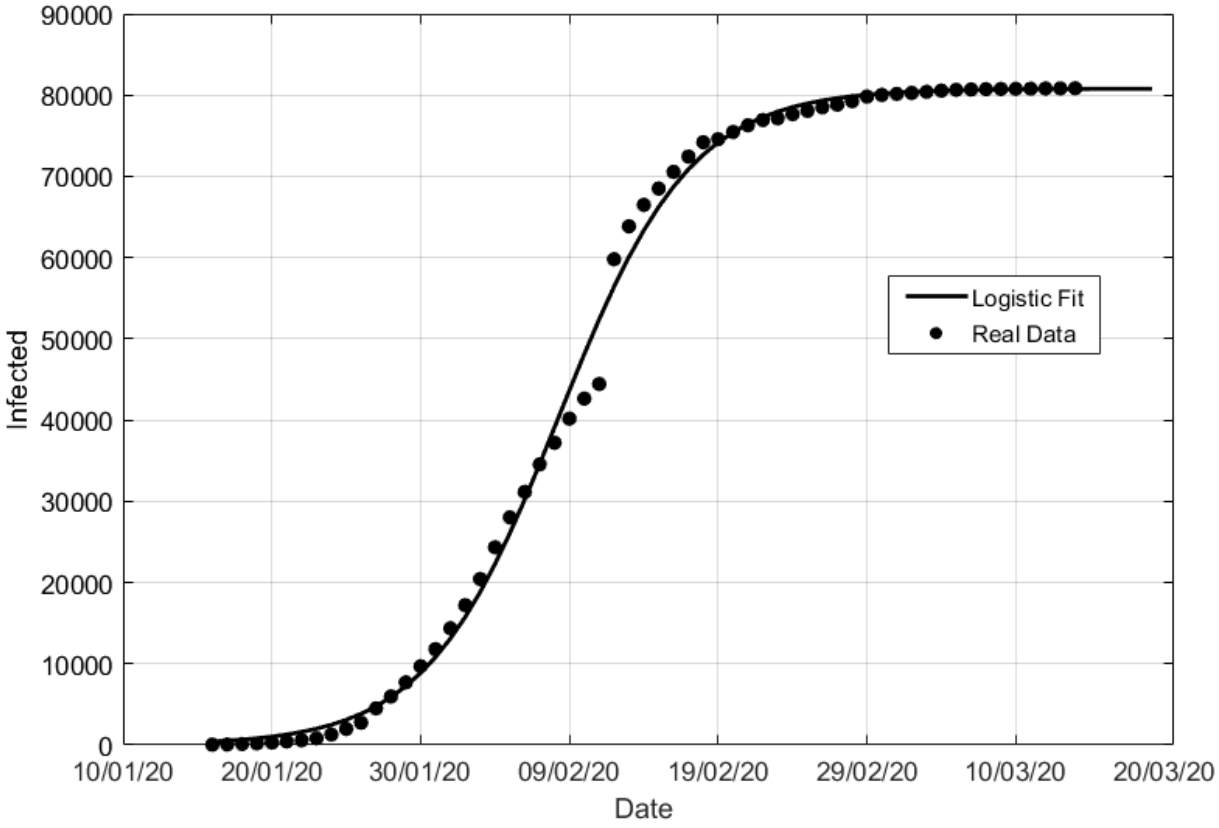

FIG. 2. Logistic (Eq. 11 fit of the Chinese number of infections $C(t)$. The best fit parameters are $a = 80800 \pm 400$, $b = 0.225 \pm 0.005$, $c = 190 \pm 25$.

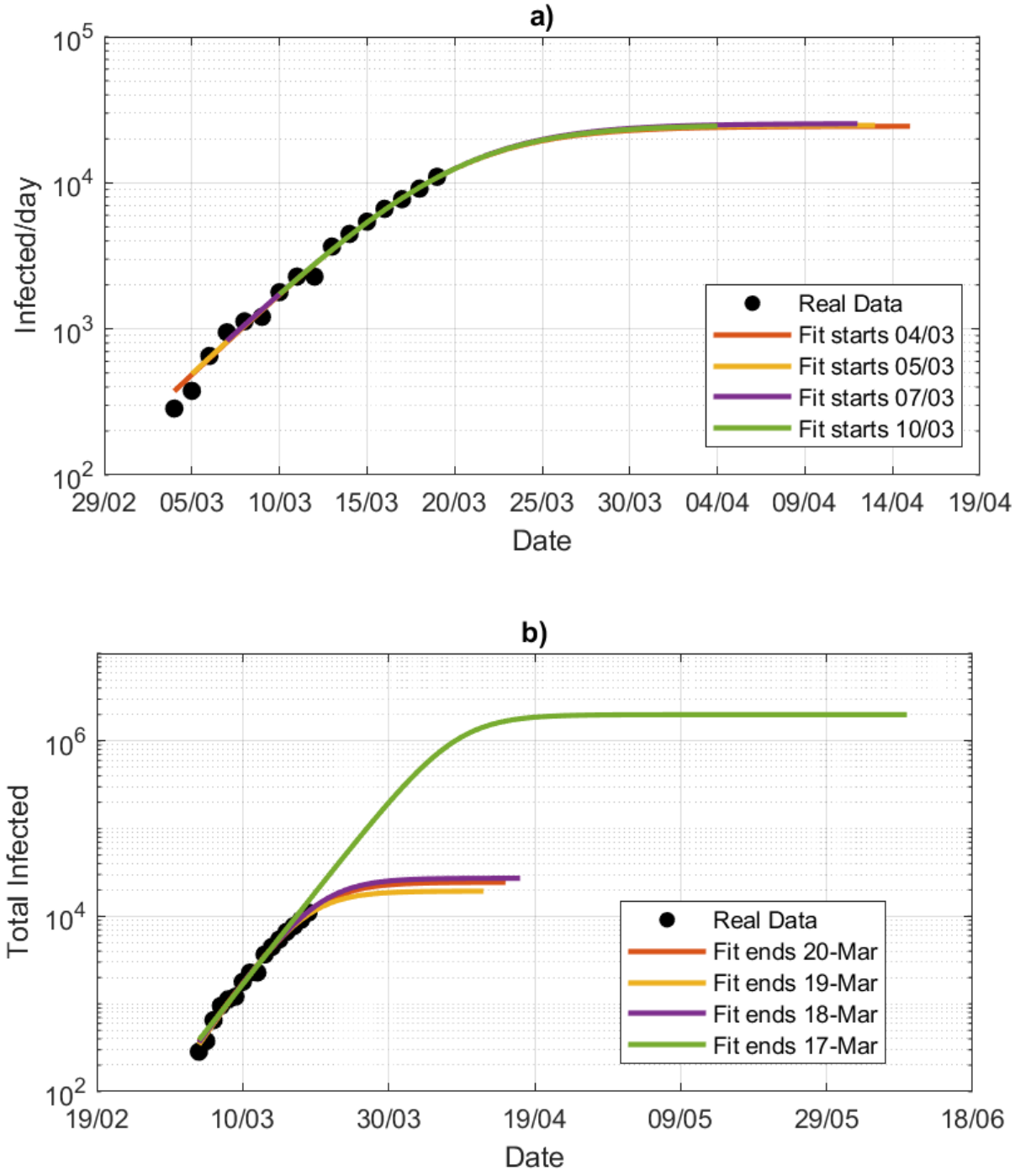

FIG. 3. Logistic distribution fits for the early stages of the epidemic in France. a) Logistic fits with data starting from different dates and ending Mar. 20. b) Logistic fits ending on different dates, but starting Mar. 04.

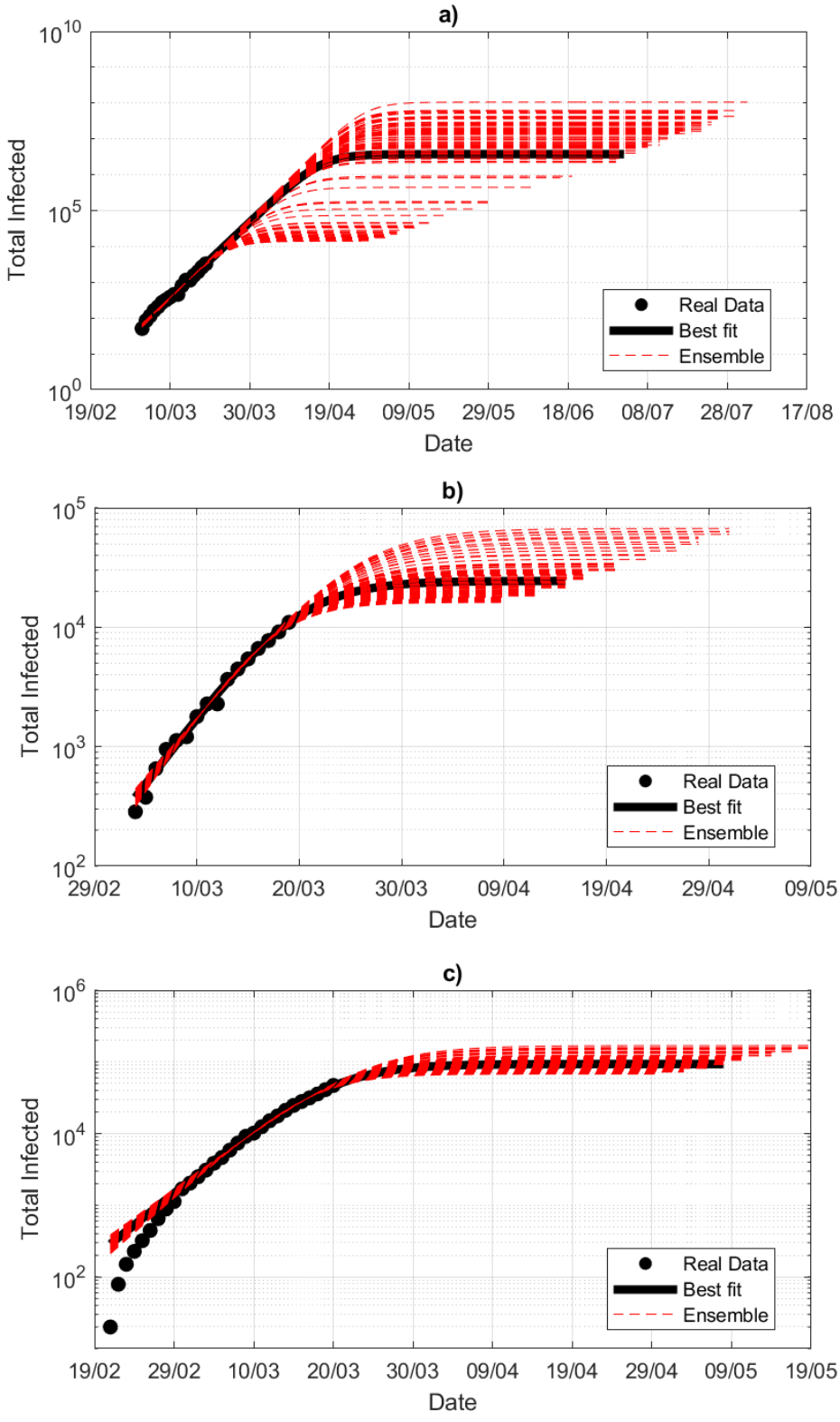

FIG. 4. Logistic distribution substituting the last data point with a random number $\xi(t^*)$ drawn from a uniform distribution with mean $C(t^*)$ and standard deviation $0.2C(t^*)$ for UK (a), France (b) and Italy (c).

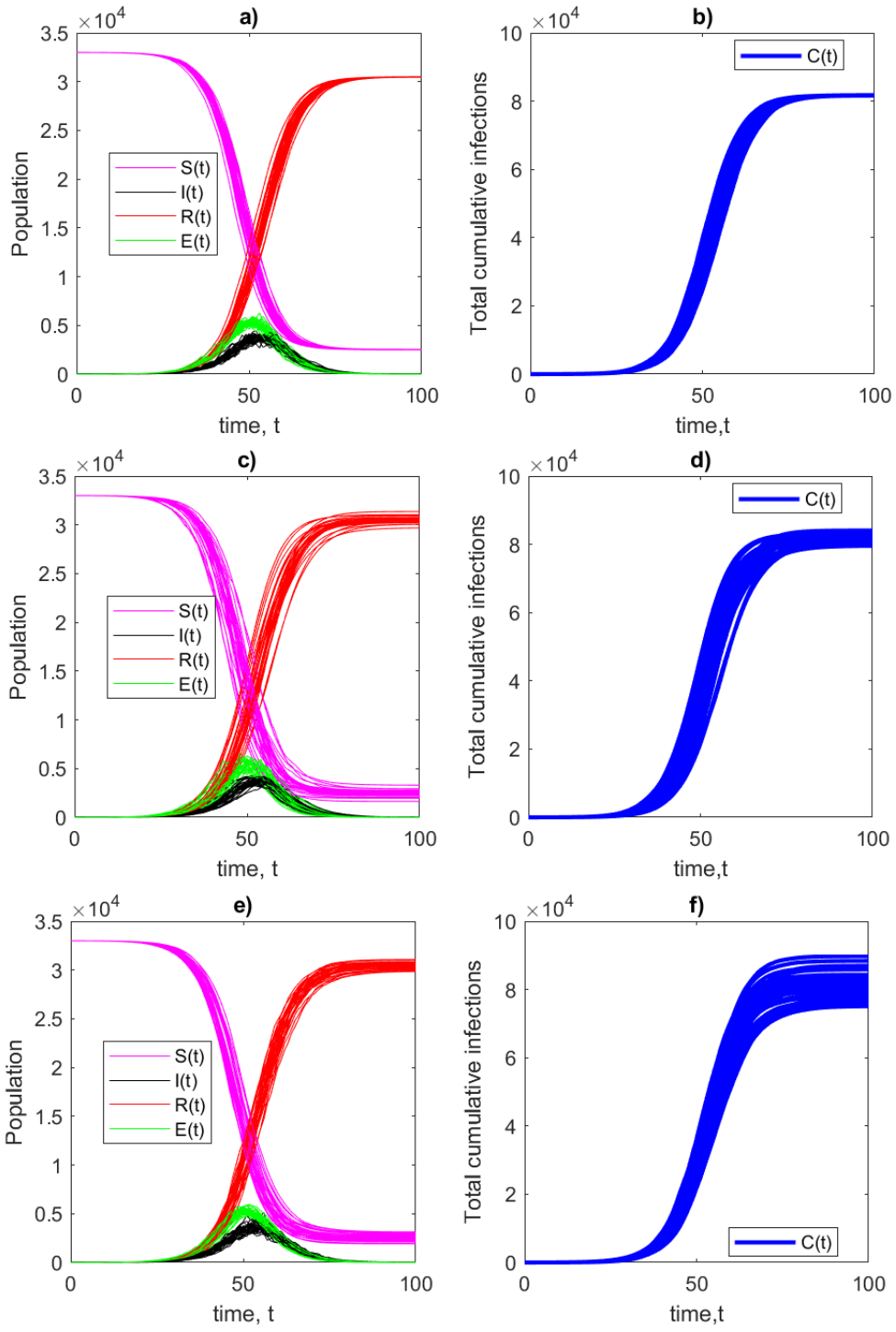

FIG. 5. Example of 30 trajectories of dynamics of stochastic Susceptible-Exposed-Infected-Recovered (SEIR) model for COVID-19, obtained replacing alternatevely α (a,b), λ (b,d) and γ with the stochastic process with Eq 12. Dynamics are integrated with a fixed initial condition and 30 noise realisations. a,c,e) Time evolution for the variables of the system, b,d,f) Time evolution for the total number of infections $C(t)$.

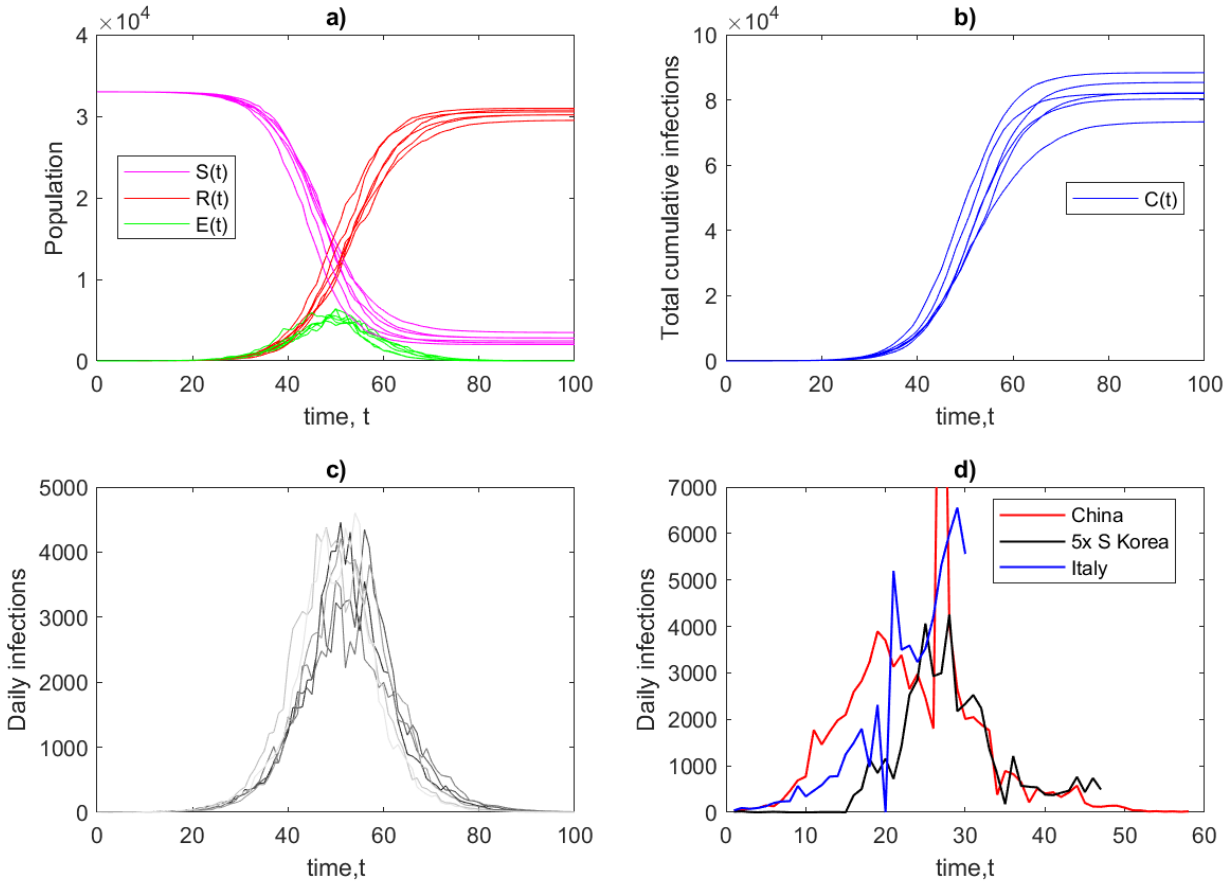

FIG. 6. Example of 6 trajectories of dynamics of stochastic Susceptible-Exposed-Infected-Recovered (SEIR) model for COVID-19, obtained replacing all parameters α, λ and γ with an independent stochastic process as in Eq 12. Dynamics are integrated with a fixed initial condition and 6 noise realisations. a) Time evolution for the variables of the system. b) Time evolution for the total number of infections $C(t)$. c) Time evolution for the daily infections. d) Comparison with daily infections in China (red, starting Dec 19. 2019), South Korea (black, starting Jan 30, 2020), Italy (blue, starting Feb 20, 2020).